PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

P-MIA: PADRÃO MÚLTIPLAS INSTÂNCIAS AUTOADAPTÁVEIS – UM PADRÃO DE DADOS PARA WORKFLOWS CIENTÍFICOS

PATRÍCIA NOGUEIRA HÜBLER

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul – PUC-RS para obtenção do título de Doutor em Ciência da Computação

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz

Porto Alegre, Dezembro de 2010.

Dados Internacionais de Catalogação na Publicação (CIP)

H878P Hübler, Patrícia Nogueira
P-MIA : padrão múltiplas instâncias autoadaptáveis - um padrão de dados para wokflows científicos / Patrícia Nogueira Hübler. – Porto Alegre, 2010.
179 f.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Informática. 2. Biologia Computacional. 3. Workflow. I. Ruiz, Duncan Dubugras Alcoba. II. Título.

CDD 005.431

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "P-MIA Padrão Múltiplas Instâncias Autoadaptáveis - Um Padrão de Dados para Workflows Científicos", apresentada por Patrícia Nogueira Hübler, como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, Sistemas de Informação, aprovada em 10/12/2010 pela Comissão Examinadora:

Prof. Dr. Duncan Dubugras Alcoba Ruiz – Orientador — PPGCC/PUCRS

Profa. Dra. Nina Edelweiss - — UFRGS

Prof. Dr. João Eduardo Ferreira - — USP

Prof. Dr. Osmar Norberto de Souza – — PPGCC/PUCRS

Homologada em 19/01/11, conforme Ata No. 001 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Rafaela, minha filha, a ti dedico este trabalho por teres sido fonte de inspiração; por teres me ensinado o verdadeiro significado da palavra amor; por teres me impulsionado a ser seu exemplo, sempre.

"É muito melhor arriscar coisas grandiosas, alcançar triunfos e glórias, mesmo expondo-se à derrota, do que formar fila com os pobres de espírito que nem gozam muito nem sofrem muito, porque vivem nessa penumbra cinzenta que não conhece vitória nem derrota."

Theodore Roosevelt

AGRADECIMENTOS

Agradecer é a forma que tenho de expressar o reconhecimento pelo envolvimento e participação de pessoas no desenvolvimento desta Tese. Pessoas que contribuíram para que este trabalho fosse finalizado. Agradeço:

- À minha família: aos meus pais, Victor Hugo e Maria Zilá, pelo exemplo em todos os dias da minha vida, ao incentivo pela continuidade dos estudos e ao apoio irrestrito em todos os sentidos; aos meus irmãos, Rafael e André, pelo carinho nas horas mais difíceis; à minha Avó Celi, pelos almoços, pela força e pelo colo; ao meu amado marido, Édison Leandro, a pessoa que escolhi para viver ao meu lado e que, durante o desenvolvimento deste trabalho: me apoiou, me incentivou, entendeu minhas ausências, e tirou do meu vocabulário a palavra *desistir*;
- Ao meu Orientador – Prof. Duncan, pelo conhecimento, pela dedicação, pelas horas de orientação, pelo incentivo (que foram muitos). Pelas palavras que evitaram a desistência... Professor Duncan: se esta Tese foi finalizada, o maior de todos os agradecimentos deve ser direcionado a ti, ao teu profissionalismo e à tua paciência – Muito Obrigada!
- Aos meus colegas da PUC: Aninha, Nelson, Márcio, Christian, Luciano, Karina, Patrícia Silveira (que mesmo longe esteve muito perto): muito obrigada pelo apoio, pelas palavras de carinho, pelas informações científicas e técnicas e, mais do que tudo isso, pela amizade!
- Aos amigos da ULBRA por terem acompanhado meus dia de loucura e terem entendido os olhos cansados e o mau humor.
- Aos mais novos amigos do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul, Campus Canoas, por participaram do fechamento deste trabalho, com palavras de incentivo e muito carinho.

Enfim, a todos que acompanharam esses anos de Doutorado. A todos que viveram, junto comigo, cada dia, cada página, cada apresentação e que me incentivaram a nunca desistir – Muito Obrigada! Agradeço a Deus por ter me concedido as forças necessárias para a finalização deste trabalho e por ter colocado, na minha vida, pessoas tão especiais!

P-MIA: PADRÃO MÚLTIPLAS INSTÂNCIAS AUTOADAPTÁVEIS – UM PADRÃO DE DADOS PARA WORKFLOWS CIENTÍFICOS

RESUMO

A busca de soluções informatizadas, com o objetivo de se obter agilidade e confiabilidade nas informações, faz com que profissionais de diferentes áreas utilizem tecnologias com propósitos semelhantes. A utilização de sistemas de gerenciamento de *workflow* é um exemplo desse tipo de solução, a qual empresas e cientistas utilizam para documentar as etapas executadas e otimizar o tempo de execução. Esta Tese apresenta um padrão capaz de manipular grandes volumes de dados e otimizar seu processamento, identificando grupos de dados promissores, como um componente de *workflows* científicos. A área de aplicação é a Bioinformática, uma área multidisciplinar, que se utiliza de várias ferramentas computacionais para a realização de seus experimentos, os quais podem demorar anos para serem finalizados. A solução proposta beneficia, dentro da Bioinformática, o desenho racional de fármacos. Assim, a contextualização da área de estudo é realizada, e é proposta uma solução para o problema por meio da definição de um padrão de dados que permite a autoadaptação de instâncias de *workflow* em execução. O P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis, assim denominado por manipular um grande conjunto de dados e por, em tempo de execução, definir as ações a serem executadas sobre os dados, é formalizado com base nas definições de redes de Petri e sua representação gráfica feita por meio de redes de Petri coloridas. Sobre o padrão, são realizados testes experimentais, os quais comprovam que, com a utilização do P-MIA, é possível reduzir a quantidade de experimentos, mantendo um critério de qualidade aceitável.

Palavras-Chave: *Workflows* Científicos, Padrão de Dados, Autoadaptação, Bioinformática

P-SaMI: SELF-ADAPTIVE MULTIPLE INSTANCES – A DATA PATTERN TO SCIENTIFIC WORKFLOWS

ABSTRACT

In the search for automated solutions, professionals of different areas use similar information technology targeting information agility and reliability. The use of a workflow management system is an example, which is employed by enterprises and scientific labs in order to record executed tasks and to optimize the elapsed time. This thesis presents a workflow pattern, as a scientific workflow component, able to manage large volumes of data and to optimize their processing, identifying promising groups into such data. Bioinformatics is our application area, a multidisciplinary area that uses a lot of computing tools for its experiments, and which can spend years to be finished. The solution proposed here benefits the rational drug design inside Bioinformatics. Then, we contextualize the area of study, and a problem solution is given through the definition of a data pattern that allows a self-adaptation of workflow instances in execution. We named P-SaMI: Self-Adaptive Multiple Instances as our proposed pattern because it is capable to manage large data sets and to take actions during processing time. P-SaMI is formally defined with Petri nets concepts and it is designed by Coloured Petri nets. We performed several tests and achieved the reduction of experiments executed, preserving an acceptable level of resulted quality.

Keywords: Scientific *Workflows*, Data Patterns, Self-Adaptive, Bioinformatics

LISTA DE FIGURAS

LISTA DE TABELAS

GLOSSÁRIO DE TERMOS E ABREVIATURAS

AMBER	*Assisted Model Building with Energy Refinement* (Construção Assistida de Modelos com Refinamento de Energia) é um programa utilizado para cálculos de proteínas e de ácidos nucléicos [SAN02]
API	*Application Programmer's Interface* - conjunto de passos ou atividades que uma aplicação usa para requerer, carregar e executar tarefas de baixo nível realizadas pelo sistema operacional (SO).
Autoadaptação	Adaptar-se sem o envolvimento de terceiros, mais especificamente, de usuários especializados.
BPM	*Business Process Management*
BPMN	*Business Process Modeling Notation* - notação visual para representação de fluxos de processos que pode ser mapeada para diversos formatos de execução
CADD	*Computer Assisted Drug Design*
Cluster	Um grupo criado com base em determinados critérios
CP-NET	*Coloured Petri Net* (Rede de Petri Colorida)
CPN	*Coloured Petri Net* (Rede de Petri Colorida)
DATA	Dado do Tipo Caractere
Desenho de Fármacos Assistido por Computador (CADD - *Computer-Assisted Drug Design*)	O planejamento de fármacos auxiliado por computador envolve todas as técnicas com o auxílio do computador usadas para descobrir, planejar e otimizar compostos biologicamente ativos com uso suposto de fármacos [SAN02]
DM	Docagem Molecular
Docagem (Docking)	Um método utilizado para detectar sítios de ligação em proteína e avaliar interações, utilizando para cálculo a energia livre de ligação entre as moléculas.
ETH	*Ethionamide*
FEB	*Free Energy of Binding*
FReDD	Flexible Receptor Docking Database
GED	Gerenciamento Eletrônico de Documentos
InhA	enzima 2-trans-Enoil ACP(CoA) *Reductase de Mycobacterium tuberculosis*
INT	Dado do Tipo Inteiro
LABIO	Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas
Ligante	pequena molécula
MTB	*Mycobacterium Tuberculosis*

NADH	*Nicotinamida Adenina Dinucleotídeo*, forma reduzida
PDB	*Protein Data Bank PDB* – É uma coleção de estruturas 3D de proteínas determinadas principalmente por cristalografia de raios X e ressonância magnética nuclear.
PIF	*Pentacyano(isoniazid)ferrate II*
PJC	Padrão Junção Combinada
P-MIA	Padrão Múltiplas Instâncias Autoadaptáveis
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul
rdP	Redes de Petri
Receptor	Um receptor é uma proteína ou um complexo de proteínas localizado no interior ou na superfície de uma célula, que reconhece especificamente e interage com um composto que atua como um mensageiro molecular (neurotransmissor, hormônio, fármaco etc.). Em um sentido mais amplo, o termo receptor é freqüentemente usado como sinônimo para qualquer sítio específico de ligação de fármacos. [SAN02]
Snapshot	Pode ser considerado como uma fotografia de um determinado objeto em um dado momento. Neste contexto, cada *snapshot* contém informações sobre as características de uma conformação, formadas por um conjunto de coordenadas X, Y, Z para cada átomo, em Angstrons (Å), além de informações topológicas referentes a cada uma.
SGWf	Sistema de Gerenciamento de *Workflow*
SWfMS	*Scientific Workflow Management Systems* (Sistemas Gerenciadores de *Workflows* Científicos)
TCL	*Triclosan*
UML	*Unified Modeling Language*
WCP	*Workflow Control-Flow Patterns*
WfMC	*Workflow Management Coalition*

SUMÁRIO

1 INTRODUÇÃO

1.1 Caracterização do Problema

Com o passar dos anos e com a evolução da computação, empresas e cientistas convergem na busca de soluções automatizadas que forneçam agilidade e confiabilidade às informações. A utilização de tecnologias de controle é amplamente difundida no meio empresarial e a implementação de sistemas de gerenciamento de *workflow* [WOR99, AAL03] propicia esse controle, automatizando etapas de um processo antes manual, diferenciando empresas e conferindo competitividade.

Com a grande aplicabilidade desses sistemas no meio empresarial, a área científica busca sua utilização, também com o objetivo de controle, mas principalmente com o objetivo de documentar as etapas executadas e de otimizar o tempo de execução. Assim, a Bioinformática, uma área multidisciplinar, na qual várias ferramentas computacionais são aplicadas para a realização de experimentos e esses experimentos repetidos por diferentes cientistas, busca, com a utilização de sistemas de gerenciamento de *workflow*, o registro das etapas executadas (proveniência) e a realização desses experimentos de forma mais rápida, uma vez que experimentos *in-silico* podem demorar anos para serem finalizados, conforme apresentado no Capítulo 3 desta Tese.

Autores como Coutinho et al. [COU10] e Mattoso et al. [MAT08] afirmam que uma das maiores características da área de Bioinformática é a manipulação de um grande volume de dados e a realização de experimentos por meio de simulação computacional, ou seja, experimentos *in silico,* demandando grande capacidade de processamento por parte dos computadores. Assim, com o avanço da biologia molecular e das ferramentas de simulação *in silico*, o planejamento de medicamentos passou a ser realizado de maneira mais lógica [MAC07a], sendo chamado de Desenho Racional de Fármacos [KUN92]. Uma das principais etapas do desenho racional de fármacos é a docagem molecular, na qual se investiga e avalia o melhor encaixe de um ligante na estrutura alvo ou receptor. A análise, mesmo que *in silico*, da interação de dados dos possíveis compostos (ligantes) com uma determinada proteína-alvo (receptor) e sua respectiva

conformação (também chamada de *snapshot*), se torna inviável de ser executada, conforme detalhado também no Capítulo 3, pois se estima que seriam necessários aproximadamente 62 trilhões de minutos (aproximadamente 117 mil anos) até o término da execução de todos os experimentos, considerando-se um tempo mínimo por execução de 1 minuto.

Portanto, além de uma contextualização breve da área, este documento também apresenta a definição de uma solução que possibilita a execução de experimentos em Bioinformática selecionados dinamicamente. Sabe-se que um dos grandes desafios da docagem molecular, além de manipular grandes volumes de dados, é a otimização de recursos computacionais. Para essa otimização, neste trabalho, utilizar-se-á o conceito de adaptação, que neste caso é aplicado aos dados em execução. Portanto, a definição de um padrão de dados que possibilite a execução de experimentos, selecionados de maneira inteligente, a partir de critérios de qualidade previamente estabelecidos, possibilita a redução do tempo total de execução, com resultados satisfatórios. Para isso, utiliza-se da definição de quais dados em execução devem ser finalizados, quais devem ser descartados e quais devem ter suas prioridades alteradas, em tempo de execução.

A principal contribuição desta pesquisa está na definição de um novo padrão de dados que aplica técnicas de adaptação dinâmica sobre um grande volume de dados, sua formalização e aplicação com dados reais, verificando o ganho obtido. Considera-se, para isso, que o ganho é a redução do número de experimentos realizados, bem como o tempo total para finalização de todo o experimento. Isso é possível ao se fazer uso dos *snapshots* que provavelmente apresentarão os melhores resultados e ao considerar a possibilidade de execução em paralelo desses experimentos.

1.2 Questão de Pesquisa

A partir da necessidade de processamento de grandes volumes de dados da área de Bioinformática, remete-se à seguinte questão de pesquisa: "Como reduzir a quantidade de *snapshots* a serem processados, reduzindo o tempo total de processamento e procurando manter o mesmo nível de acerto na identificação de compostos promissores?”

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desta Tese de Doutorado é melhorar o tempo final de processamento, reduzindo a quantidade de experimentos de docagem molecular, com base nos resultados obtidos em tempo de execução.

1.3.2 Objetivos Específicos

Como objetivos específicos buscam-se:

- definir um padrão de dados capaz de ser utilizado pela área de Bioinformática e por outras áreas que apresentem características semelhantes;
- definir uma função que não descarte dados que apresentem a probabilidade de serem promissores;
- reduzir a quantidade total de dados a serem processados;
- buscar manter a qualidade dos dados processados.

1.4 Método de Pesquisa

Para o desenvolvimento desta Tese foram realizados: um levantamento, com o intuito de identificar as necessidades de processamento do LABIO - PUCRS (Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas da Pontifícia Universidade Católica do Rio Grande do Sul); uma revisão da literatura para identificar aspectos vistos como importantes nos trabalhos seminais da área de automação de *workflows* e seus padrões; a definição de uma proposta de modelo de dados alinhado com a literatura identificada e com o estado da arte apresentado; e, por fim, uma pesquisa experimental, a partir da determinação de um objeto de estudo, selecionando as variáveis que seriam capazes de influenciá-lo e de um experimento com dados reais para a verificação e análise do modelo proposto.

1.5 Organização do Trabalho

A Tese está organizada da seguinte forma:

- No Capítulo 2 são apresentados conceitos sobre *workflows* científicos e *workflows* de negócios, bem como modelos utilizados para a formalização de processos: rede de Petri e redes de Petri Coloridas. Esses conceitos são fundamentais para o entendimento do padrão definido nesta Tese.

- No Capítulo 3 a Bioinformática, uma área de conhecimento convergente e multidisciplinar, área com o qual o padrão apresentado nesta Tese é validado, é apresentada. Além da apresentação de forma genérica da área, um detalhamento sobre algumas pesquisas realizadas pelo LABIO, que são convergentes ao trabalho desenvolvido nesta Tese, também é encontrado.
- No Capítulo 4 é apresentada uma série de padrões de dados, os quais objetivam capturar as diferentes formas de representação e utilização dos dados sobre *workflows*, além de alguns padrões de fluxo, utilizados como base para o desenvolvimento do trabalho aqui desenvolvido. O estudo destes padrões objetiva a identificação da existência de algum que seja capaz de atender à necessidade de áreas como a Bioinformática: manipular grandes volumes de dados, com características semelhantes, na menor quantidade de tempo possível.
- O Capítulo 5 contém a formalização do padrão definido nesta Tese, identificando suas principais características. Além disso, apresenta os padrões de dados utilizados pelo padrão, bem como sua representação gráfica por meio de redes de Petri coloridas.
- O funcionamento do padrão definido nesta Tese é apresentado no Capítulo 6, identificando suas características e regras para implementação. Além disso, a integração com as pesquisas realizadas no LABIO, por meio da substituição de etapas do *workflow* científico, desenvolvido por Karina Machado [MAC07a], também é detalhada.
- Para a validação do funcionamento do padrão proposto, o Capítulo 7 contém resultados de testes que foram realizados com grupos definidos a partir de uma função de similaridade. Os testes apresentados foram realizados com dois ligantes: PIF [OLI04] e NADH [DES95].
- No Capítulo 8 os trabalhos relacionados com o trabalho desenvolvido são apresentados e, no final deste capítulo, uma tabela de comparação, contendo as principais características dos diferentes trabalhos sintetiza os estudos realizados.
- As considerações finais do trabalho desenvolvido nesta Tese de Doutorado, as principais contribuições e os trabalhos futuros, são apresentados na sequência.

2 REVISÃO DA LITERATURA

Este capítulo apresenta conceitos sobre *workflows* de negócio e *workflows* científicos, suas características, semelhanças e diferenças, subsidiando o desenvolvimento do padrão definido nesta Tese, cuja comparação entre esses dois tipos é realizada nas considerações finais deste capítulo. Além dos dois tipos de *workflows*, este capítulo também apresenta redes de Petri e redes de Petri coloridas, utilizadas nesta Tese para a formalização de processos e para a representação do padrão definido, podendo este último ser encontrado no capítulo 5.

2.1 *Workflows* de Negócios

O bom funcionamento de uma empresa depende, essencialmente, das possibilidades de acesso à informação e do processo decisório baseado sobre elas. O trabalho administrativo é um trabalho em equipe, onde cada membro, um ator, tem um conjunto próprio de atribuições, responsabilidades e autonomia para a realização de suas atividades. Essas atividades, contudo, nem sempre são independentes. Geralmente, há uma interdependência na realização das atividades, causada pela manipulação e transformação de elementos de informação compartilhados em instantes diferentes. Há, em consequência, uma estreita cooperação no trabalho realizado.

Para organizar e assegurar a qualidade dessa cooperação é importante que existam modelos com capacidade de descrição e ambientes de automação computacional, possibilitando o emprego desses tipos de descrição para o suporte computacional do fluxo de trabalho. Uma das soluções para atender a esses requisitos é a utilização da tecnologia de *workflow.* Essa tecnologia permite representar, manipular e monitorar informações relativas ao fluxo de trabalho e as utiliza para gerenciar, coordenar e controlar o trabalho administrativo de maneira mais eficiente. Esse tipo de suporte computacional é chamado automação de *workflow*.

Para a WfMC (*Workflow Management Coalition* – [WOR99]) *workflow* é a automação do processo de negócio, na sua totalidade ou em partes, onde documentos, informações ou tarefas são passadas de um participante para o outro para execução de uma ação, de acordo com um

conjunto de regras de procedimentos. Plesums [PLE02] afirma que *workflow* é tradicionalmente definido em relação a termos utilizados em escritórios, movimentando documentos, processando ordens e executando chamadas. Entretanto, o autor também afirma que esse princípio é o mesmo utilizado para outros sistemas.

Segundo Plesums [PLE03], sistemas de automação de *workflow* tornaram-se populares inicialmente com a utilização de sistemas de gerenciamento de imagens (*Document Imaging* – GED: Gerenciamento Eletrônico de Documentos). O autor ainda afirma que o gerenciamento automatizado de *workflow* tem sido continuamente discutido nos últimos 20 anos e que muitas pessoas o visualizam como parte de um sistema de manipulação de imagens. Em [PRI03], Prior relata que, na década de 90, os *workflows* eram utilizados como solução para a reengenharia de processos de negócio, buscando, simplesmente, automatizá-los. O foco estava, apenas, na utilização da tecnologia, ou seja, nas aplicações e em sistemas com baixa interação humana e isso não significava garantia de sucesso.

A automatização de sistemas de *workflow* é realizada por meio de um Sistema de Gerenciamento de *Workflow* (SGWf) que, conforme a WfMC [WOR99] e Aalst [AAL03], é um sistema que define, cria e gerencia a execução de *workflows* por meio do uso de software, executando em um ou mais servidores. Sistema esse, apto a interpretar a definição dos processos, a interagir com seus participantes e, quando requerido, ativar ferramentas de software e aplicativos.

Com a habilidade de se modelar processos de negócios, monitorá-los em tempo real e, por esses processos serem mais simples de serem controlados, cresceu o interesse pelo gerenciamento dos processos de negócio (*Business Process Management* – BPM). Para Aalst [AAL03], BPM é definido como o "*apoio aos processos de negócio, usando métodos, técnicas e software para projetar, desempenhar um papel, controlar e analisar processos operacionais, envolvendo pessoas, organizações, aplicações, documentos e outras fontes de informação*".

As principais funcionalidades oferecidas por um SGWf são definidas por Georgakopoulos [GEO95], Hollingsworth [HOL95] e Leymann [LEY00]:

- a definição e modelagem dos processos de *workflow* e suas respectivas atividades;
- o gerenciamento dos processos de *workflow* em um ambiente operacional e o ordenamento das várias atividades a serem manipuladas como parte de cada processo; e

- o controle das interações entre usuários e aplicativos no processamento dos vários passos das atividades.

A Figura 1 apresenta a relação entre os principais termos envolvidos na automatização de processos (*workflows*). O vocabulário utilizado nesse trabalho é o definido pela WfMC. Conforme apresenta a Figura 1, um processo de negócio é elaborado por meio da sua definição, ou seja, o que deve acontecer quando da execução desse processo, e é gerenciado por um sistema específico. Esse sistema deve controlar os aspectos automatizados. A definição dos processos é composta por atividades que, em um sistema de *workflow*, correspondem a uma etapa a ser executada dentro de um processo, podendo ser executadas de forma manual ou automatizada. As atividades automatizadas são, quando gerenciadas pelo SGWf, armazenadas como instâncias das atividades. A cada implementação de um novo processo informatizado, suas atividades devem ser analisadas e definidas.

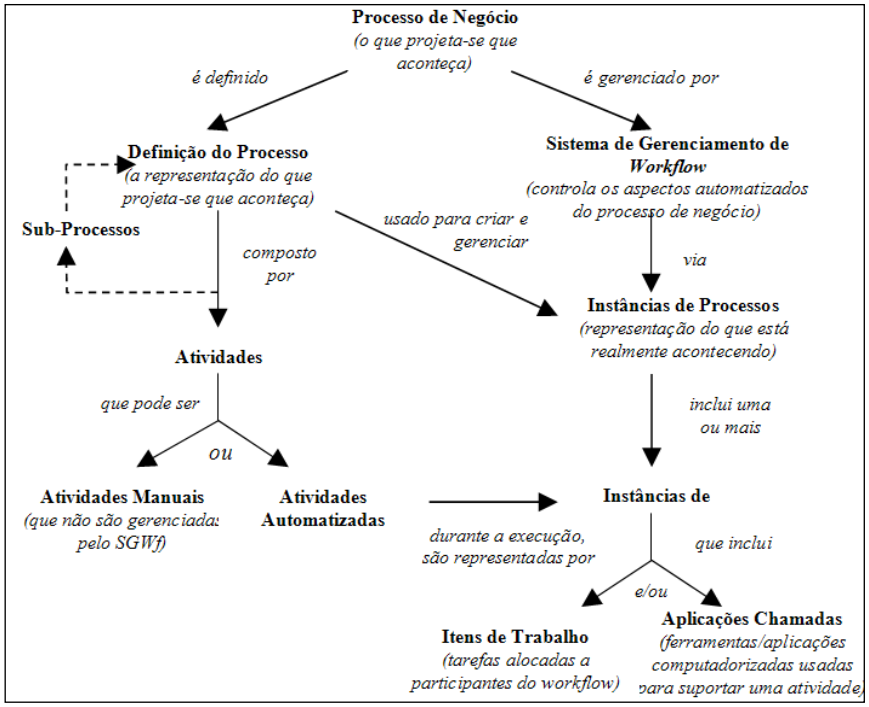

Figura 1 – Principais termos envolvidos na automatização de processos [WOR99]

2.1.1 Tipos de *Workflow*

Na literatura são encontradas diferentes classificações para *workflow*. A mais popular é a de Georgakopoulos [GEO95], que os classifica como: *ad-hoc*, administrativo e de produção. A classificação feita por Leymann [LEY00] acrescenta *workflows* colaborativos:

- *Workflow Ad-Hoc*: *Workflows ad-hoc* executam processos de negócios, tais como documentação de produtos ou venda de produtos, onde não há um padrão pré-determinado de movimentação de informação entre pessoas. Tarefas do tipo *ad-*

hoc envolvem a coordenação humana ou a co-decisão, conforme Schael et al. [SCH91]. A ordenação e a coordenação de tarefas em um *workflow* do tipo *ad-hoc* são controladas por usuários, ou seja, não são automatizadas. Essa classe de *workflow* envolve pequenos grupos de profissionais que têm a intenção de apoiar pequenas atividades que requerem uma solução rápida. Leymann [LEY00] afirma que esse tipo de *workflow* é aplicado a processos que apresentem baixo valor para o negócio e uma baixa taxa de repetição.

- *Workflow* Administrativo: Sistemas de *workflow* administrativos destinam-se a processos simples e estruturados. São processos burocráticos, repetitivos, com regras bem definidas e do conhecimento de todos os participantes do fluxo [GEO95]. Da mesma forma que os *ad-hoc*, também apresentam baixo valor para o negócio [LEY00].
- *Workflow* de Produção: Um *workflow* de produção envolve processos de negócios repetitivos e previsíveis. Englobam processamentos complexos, com acesso a múltiplos sistemas de informação. São caracterizados por fornecerem muito valor ao negócio [LEY00].
- *Workflow* Colaborativo: Os sistemas de *workflow* colaborativo são adequados para processos que envolvam trabalho cooperativo realizado por equipes de pessoas com objetivos comuns, como os sistemas de *groupware* que, conforme Aalst [AAL01], são sistemas que suportam esse estilo de trabalho. Podem ser adotados para automatizar processos empresariais críticos que não são orientados à transação. Esses processos podem ser executados poucas vezes. Entretanto, sua execução e seu sucesso são essenciais para a Organização [LEY00].

2.2 *Workflows* Científicos

Deelman e Gil [DEE09] afirmam que nas últimas duas décadas a revolução está sendo conduzida pela ciência e pela engenharia e que supercomputadores têm sido utilizados para simularem sistemas que envolvem dados complexos e visualização das etapas desses processamentos. Apresentam como um cenário típico a necessidade cíclica de envio de dados para supercomputadores para análise ou simulação, com armazenamento dos resultados obtidos. Para os autores, sistemas que implementam *workflows* científicos objetivam automatizar esse ciclo, possibilitando aos cientistas foco em suas pesquisas e não no gerenciamento computacional. Assim, definem um *workflow* como se referindo à sequência de atividades necessárias para se

gerenciar um processo de negócio, um processo científico ou um processo de engenharia. Uma instância desse *workflow* é a instanciação de um problema particular e inclui a definição de dados de entrada. O objetivo de sistemas de *Workflows* Científicos é, portanto, possibilitar um ambiente de programação especializada para simplificar os esforços investidos por cientistas na finalização de um experimento científico com suporte computacional [TAY06].

Conforme Ludaescher et al. [LUD06, LUD09] o termo "*workflow*" tem sido tradicionalmente utilizado no contexto de aplicações de negócios. Para os autores um *workflow* científico facilita ou automatiza um "processo científico", por exemplo, a execução de um experimento em uma ferramenta, seguido pelo pós-processamento dos dados e a interpretação desses resultados. Definem um *workflow* científico como uma descrição dos processos que um cientista necessita executar, e na ordem correta de execução, para criar um produto para a ciência.

2.2.1 Perspectivas de Usuários

Ludaescher et al. [LUD06, LUD09] apresentam as diferentes perspectivas de usuários em relação a *workflows* científicos:

- *Visão de Cientista do Domínio*: do ponto de vista de um usuário final, a automação é o maior benefício da abordagem de *workflows* científicos. Por exemplo, um cientista executa diferentes conjuntos de dados com a mesma sequência de etapas, usando ou não diferentes parâmetros de configuração. O sistema está apto a registrar todas as interações dos usuários durante a execução do *workflow*, além de registrar informações que facilitem a interpretação dos dados registrados e um acompanhamento em caso de necessidade. Outros requisitos da perspectiva dos cientistas incluem: o projeto de *workflow* deve ser intuitivo e direcionado à reutilização e reconfiguração. Além disso, deve possibilitar a execução por longos períodos de tempo monitorando-a e, se necessário, poder suspender ou abortar essa execução de forma remota ou dinamicamente alterar seus parâmetros de configuração. A especificação de um *workflow* científico deve ser compartilhada e discutida com a comunidade científica.

- *Visão de Engenheiro do Workflow*: os cientistas do domínio são os usuários finais e participam cada vez mais dos projetos dos *workflows* científicos. O papel de engenheiro do *workflow* é similar ao de engenheiro de software: conhecer as diferentes ferramentas disponíveis que podem ser utilizadas por sistemas de

workflows científicos. Muitas ferramentas, por exemplo, podem ser executadas por linhas de comando, outras por *web services*. Enquanto cientistas da computação desenvolvem e aprimoram sofisticados métodos, alguns detalhes podem estar omissos aos engenheiros de *workflow*, podendo utilizar uma linguagem de modelagem de alto nível. Outro objetivo dos engenheiros de *workflow* é o desempenho: a partir de uma descrição de *workflow*, analisar como o *workflow* poderia ser executado eficientemente em um ambiente de *cluster* ou grade, ou quais as vantagens das tarefas serem executadas em *pipeline* ou com a utilização de paralelismo.

- *Visão do Cientista da Computação*: a linha que separa os papéis e visões de engenheiro de *workflow* e cientista da computação não é muito clara. Por exemplo, *workflows* científicos modelam e projetam métodos que podem ser baseados em combinações de técnicas de engenharia de software, teorias de bancos de dados, otimização de consultas, processamento de fluxo, programação funcional. Em particular, quando se cria um novo paradigma de modelagem, um cientista da computação deve estar interessado nos formalismos que possibilitam a análise estática dos *workflows* para garantir algumas propriedades como definição de tipos, liberdade de *deadlocks*, entre outros. Um modelo formal de computação auxilia cientistas da computação a definirem estratégias de planejamento e otimização de *workflows*. Alguns problemas relacionados a *workflows* não são simples e requerem heurísticas que garantam um planejamento eficiente. Outro desafio da ciência da computação é projetar e implementar sistemas de gerenciamento de proveniência, eficientemente, para *workflows* científicos.

Ludaescher et al. [LUD06] e [LUD09] afirmam que muitas pesquisas sobre *workflows* científicos estão focadas na otimização de algoritmos, nos desempenhos dos sistemas e nos requisitos de memória. Entretanto, definem que o recurso mais precioso em *e-Science* é o tempo humano.

2.2.2 Níveis e Abordagens para *Workflows* Científicos

Para Rodriguez et al. [ROD08] assim como *workflows* de negócios, especificações de *workflows* científicos podem ser executadas por softwares apropriados para execuções científicas, como os Sistemas de Gerenciamento de *Workflows*. Para os autores, o significado dos *links*

(conexões entre as tarefas) em *workflows* científicos pode ser analisado sob diferentes aspectos: direcionado a controle ou direcionado a dados. Algumas abordagens utilizam elementos de ambos:

- Na abordagem direcionada a *Controle*, os *links* entre as tarefas representam restrições de controle para o desempenho de cada tarefa. Muitas estruturas de controle podem ser encontradas, desde as mais básicas como sequências que representam a execução ordenada de tarefas até as mais complexas estruturas de controle como *splits, joins, loops* etc. Os padrões de *workflows* foram propostos dentro do domínio de negócios, para controlar e descrever os possíveis comportamentos do controle. [ROD08]
- Na abordagem direcionada a *dados*, os *links* entre as tarefas representam dependências de dados. Uma tarefa consome dados e produz dados. Cada tarefa pode ser iniciada quando uma determinada entrada está disponível. Essa abordagem tem a vantagem de que a execução paralela de tarefas independentes é modelada de forma livre. [ROD08]

Os autores também afirmam que existem diferentes argumentos para se utilizar uma ou outra abordagem. Por um lado, para Pautasso e Alonso [PAU06], a utilização direcionada a controle fornece maior domínio sobre a ordem atual de execução das tarefas. Por outro lado, as representações direcionadas a dados são mais simples para usuários finais. A abordagem direcionada simplesmente a dados, entretanto, não é suficientemente expressiva para modelar o comportamento eventualmente iterativo dos processos. Muitas propostas seguem uma abordagem híbrida, baseada na combinação de controle e dados, utilizando as principais características de cada uma delas.

Rodriguez et al. [ROD08] também apresentam que a execução de *workflows* científicos envolve o processamento de uma especificação de *workflow*, coordenando o uso apropriado de recursos ou serviços. Apresentam a definição de três níveis de *workflows* [MCG06]:

- *Workflows Abstratos*: alto nível de abstração do *workflow* que contém a informação sobre o que é feito em cada tarefa e como as tarefas são interconectadas. Não existe menção de como é a entrada dos dados ou a implementação das tarefas. Como exemplo, um *workflow* abstrato pode ser descrito como uma equação linear que recebe entradas de um dispositivo e envia os resultados a um monitor;

- *Workflows Concretos*: o mapeamento para *workflows* concretos envolve a informação sobre a implementação de cada tarefa e os recursos a serem utilizados, informações sobre quais métodos são chamados, bem como o formato de dados necessário para troca de informações;
- *Workflows Instanciados*: um *workflow* instanciado representa o mapeamento atual do *workflow* concreto dentro de recursos computacionais, envolvendo as respectivas entradas de dados e suas respectivas saídas. Esse nível pode ser decomposto em mais dois, conforme Deelman e Gil [DEE06]:
    - *Instância de Workflow*: descreve o *workflow* para o nível de aplicação sem indicar os recursos que ele necessita para executar. Envolve o *workflow* concreto e seus dados, sem detalhes sobre a execução do sistema;
    - *Workflow executável*: é criado pelo mapeamento da instância de *workflow* dentro de recursos computacionais.

2.2.3 Estratégias de Processamento de *Workflows*

Em Glatard et al. [GLA08] encontra-se a distinção entre duas estratégias de processamento de *workflows*: baseada em tarefas e baseada em serviços. As soluções baseadas em tarefas, também denominadas de computação global, apresentam cada tarefa formalmente descrita antes de ser submetida à execução. Os usuários definem qual é a tarefa a ser executada. Os arquivos de códigos executáveis, dados de entrada e parâmetros são utilizados para habilitar a execução. O sistema de gerenciamento de *workflow* encontra o recurso computacional apropriado para a execução. As soluções baseadas em serviço, onde a execução é manipulada por um serviço externo, são invocadas por uma interface. Os serviços são vistos como caixas pretas de um sistema de gerenciamento de *workflow*, onde apenas a chamada por intermédio das interfaces é conhecida. O sistema de gerenciamento do *workflow* deve encontrar os serviços apropriados para executar as funcionalidades requeridas.

Para os autores, *workflows* baseados em tarefas e baseados em serviços diferem na manipulação dos dados. A natureza não estática da descrição dos dados na abordagem baseada em serviços habilita extensões dinâmicas de conjuntos de dados a serem processados: um *workflow* pode ser definido e executado utilizando um conjunto completo de dados de entrada que não são conhecidos em detalhes, pois novos dados podem ser construídos por novas fontes. Os autores ainda afirmam que é comum em aplicações científicas que a aquisição dos dados seja

realizada por processos “pesados” e que os segmentos de dados sejam produzidos progressivamente. Alguns *workflows*, inclusive, podem produzir seus próprios dados, parando a produção desses dados quando as entradas suficientes tiverem sido criadas para produzir os resultados necessários. Finalmente, essa dinamicidade é requerida quando o dado de entrada é o resultado de uma consulta à base de dados, na qual o tamanho da resposta não é conhecido antecipadamente. Para os autores, uma diferença significante entre as abordagens de tarefa e serviço está na habilidade dos serviços trabalharem com conjuntos de dados dinâmicos, na qual existam laços, coletando dados de diferentes fontes.

Conforme Glatard et al. [GLA08], do ponto de vista de usuário, a principal diferença entre as abordagens surge quando se considera a re-execução de uma mesma aplicação de *workflow* sobre diferentes conjuntos de entradas de dados, comumente feita pela instanciação de aplicações em paralelo. No *workflow* baseado em tarefa, uma tarefa computacional é definida por um simples conjunto de dados de entrada e um processamento simples, executando o mesmo processamento sobre dois diferentes conjuntos de dados, resultando na descrição de duas tarefas independentes. Essa abordagem enfatiza a replicação de grafos de execução para todos os dados de entrada a serem processados. Para a abordagem baseada em serviço, uma simples extensão da abordagem baseada em tarefas é proposta, na qual uma tarefa genérica pode ser descrita por um conjunto de dados de entrada, resultando na execução de múltiplos *jobs*: um por dado de entrada. Tarefas parametrizadas não podem ser utilizadas em um *workflow*, no qual cada tarefa necessita ser replicada para todos os conjuntos de dados. Por outro lado, a abordagem baseada em serviços facilmente acomoda conjuntos de entradas de dados pela flexibilidade da abordagem. Apesar disso, essas são tecnologias de momento, devendo ser definidas em momento de implementação.

2.2.3.1 Paralelismo em *Workflows* Científicos

Glatard et al. [GLA08] assumem que o primeiro nível de paralelismo que pode ser explorado é o paralelismo intrínseco de *workflows*, representado pelo esquema modelado. Essa implementação é comum e é realizada pelos sistemas de gerenciamento de *workflow*. Quando se considera aplicações com grande volume de dados, muitos conjuntos de dados de entrada são processados sobre um determinado *workflow*, beneficiando-se de um grande número de recursos disponíveis em um ambiente eventualmente paralelo, onde serviços de *workflow* podem ser instanciados como muitas tarefas computacionais, executando em diferentes recursos de hardware e processando diferentes dados de entrada em paralelo. Os autores também definem paralelismo de outros dois níveis:

- O *paralelismo de dados* denota que um serviço está habilitado a processar muitos fragmentos de dados simultaneamente com perda mínima de desempenho. Essa capacidade envolve o processamento de dados independentes em diferentes recursos computacionais. Para os autores, então, o paralelismo de dados ocorre quando diferentes conjuntos de dados aparecem em uma simples tarefa do *workflow*, enquanto paralelismo intrínseco ocorre quando o mesmo conjunto de dados aparece várias vezes em diferentes tarefas de um mesmo nível.

- Conjuntos de dados de entrada são independentes entre si, para serem instanciados em paralelo. O *paralelismo de serviço* denota que o processamento de dois diferentes conjuntos de dados de entrada por dois serviços são totalmente independentes. Esse modelo em *pipeline* pode ser adaptado por execuções sequenciais em *workflows* baseados em serviços. O paralelismo de serviço ocorre quando diferentes conjuntos de dados aparecem em diferentes células de um mesmo nível. Os autores supõem que cada serviço processe um único conjunto de dados ao mesmo tempo, fazendo, assim, paralelismo de serviços.

2.3 Redes de Petri: Um Formalismo Utilizado para Representação de *Workflows*

Esta seção apresenta Redes de Petri e Redes de Petri Coloridas, identificando suas formas de representação, analisando suas características e particularidades.

2.3.1 Redes de Petri (rdP)

Historicamente, as Redes de Petri surgiram de um estudo realizado por Carl Adam Petri, em sua Tese de Doutorado em 1962 [PET62] Apud [AAL98]. Desde então, a utilização e o estudo dessas redes têm crescido consideravelmente, o que faz com que seja um dos exemplos mais conhecidos de teoria de ordem parcial para modelagem e análise de sistemas concorrentes. De acordo com Braghetto [BRA06], a popularidade das redes de Petri (rdP) como ferramenta para estudo de sistemas se deve à sua representação gráfica de fácil compreensão e ao seu potencial matemático para a análise de processos. Uma das grandes vantagens dessas redes é a sua flexibilidade e o alto poder de abstração, possibilitando introduzir e adaptar elementos gráficos e/ou matemáticos com a finalidade de aproximar usuários de sistemas mais específicos. Heuser [HEU88] afirma que as Redes de Petri por si só não têm significado algum e que para utilizá-las é necessário que elas tenham uma interpretação. Além disso, muitas vezes, as redes ainda são complementadas com textos, chamados de anotações.

Uma rede de Petri clássica é um grafo bipartido dirigido (no qual o conjunto de vértices V pode ser dividido em dois conjuntos V1 e V2 disjuntos, sendo que cada aresta do grafo conecta um vértice pertencente a V1 a um vértice pertencente a V2 ou vice-versa) com dois tipos de nodos, chamados: lugares e conexões (transições) [HEU88, AAL02]. Cada nodo é conectado por arcos (direcionados) orientados e as ligações com dois arcos de um mesmo tipo não são permitidas. Os lugares são representados tipicamente por círculos e as transições por retângulos, conforme Figura 2.

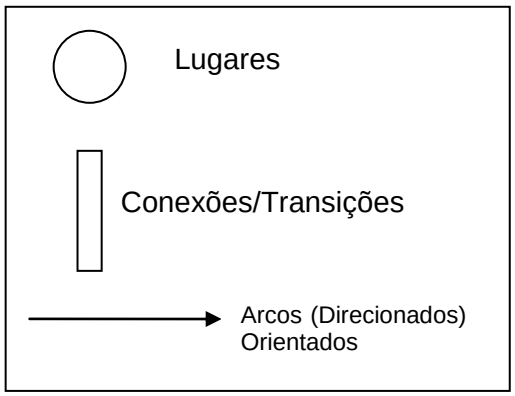

Figura 2 – Elementos básicos das redes de Petri

Os conceitos associados a cada um dos itens de uma rdP são [BRA06, AAL02, HEU88]:

- *Lugar* (representado graficamente por um círculo): modela uma condição que deve ser satisfeita para que o disparo da ação seja realizado. Conforme Braghetto [BRA06], "um lugar de entrada de uma transição geralmente expressa uma pré-condição, um dado/sinal de entrada, um recurso de hardware/software requerido ou ainda um *buffer*. Um lugar de saída pode ser compreendido como uma pós-condição, um dado/sinal de saída, um recurso liberado, uma conclusão ou um *buffer*."

- *Conexão/Transição* (representada graficamente por um retângulo ou barra): pode ser compreendida como uma ação/evento, ou um processamento de um sinal, ou ainda uma tarefa;

- *Arco (direcionado) orientado*: liga um lugar a uma transição ou vice-versa, encadeando condições e eventos;

- *Marca ou ficha* (*tokens*): representa um recurso disponível. O posicionamento dessas fichas nos lugares do grafo constitui a marcação da rdP. Cada lugar pode possuir 0 (zero) ou mais fichas. A evolução da marcação permite modelar o

comportamento dinâmico do sistema. Essas marcas são representadas na forma de pontos dentro dos lugares;

- *Peso*: cada arco possui um peso associado a ele. O peso indica quantas marcas uma transição consome de um lugar de entrada ou quantas marcas uma transição acrescenta em um lugar de saída. Quando um arco não possui um peso explicitamente indicado no grafo, considera-se que o seu peso é 1.

O grafo de uma rdP é orientado e seus arcos possuem pesos. Uma rede de Petri em que todos os arcos possuem peso 1 é classificada como ordinária. O disparo das transições (execução das ações) é controlado tanto pelo número de marcas quanto pela sua distribuição nos lugares. Uma transição *t* está habilitada se, e somente se, todos os lugares de entrada ($p_i \in P$) de *t* são marcados com pelo menos (p, t) *tokens*, onde w(p,t) é o peso do arco de p para t. Formalmente, a rede de Petri é dada por uma quíntupla, RP = <P, T, F, W, M_0>, em que [MUR89]:

- $P = \{p_1, p_2, ..., p_m\}$ conjunto finito de lugares;
- $T = \{t_1, t_2, ..., t_n\}$ conjunto finito de transições;
- $F \subseteq (P \times T) \cup (T \times P)$ conjunto de arcos;
- $W: F \rightarrow \{1, 2, 3, ...\}$ função de pesos dos arcos;
- $M_o: P \rightarrow \{0,1, 2, 3, ...\}$ marcação inicial da rede;
- $P \cap T = \varnothing$ e $P \cup T \neq \varnothing$.

Denota-se por N uma rede de Petri com estrutura N = {P, T, F, W} sem qualquer marcação inicial especificada. Denota-se por (N, M_0) uma rede de Petri com marcação inicial. Já para Aalst [AAL98] as Redes de Petri são uma tripla <P, T, F> onde:

- P é um conjunto finito de lugares;
- T é um conjunto finito de transições (conexões) com $P \cap T = \varnothing$;
- $F \subseteq (P \times T) \cup (T \times P)$ é um conjunto de arcos (fluxos).

A definição de Aalst [AAL98] é mais simples que a de Murata [MUR89], pois não possui uma marcação inicial da rede. Considera que um lugar p é chamado de lugar de entrada de uma transição se existir um arco direcionado de p para t. Um lugar p é chamado de lugar de saída de uma transição se existir um arco direcionado de t para p. Além disso, Aalst [AAL98] restringe os arcos com peso 1 e afirma que, no contexto de *workflows,* não faz sentido ter outros pesos, pois os lugares correspondem a condições.

Existe uma série de métodos que permitem analisar um grande número de propriedades de sistemas descritos por rdP. As propriedades das redes de Petri podem ser divididas em dois grandes grupos: as comportamentais, que dependem da marcação inicial e as estruturais, que não dependem.

A Figura 3 apresenta uma modelagem de processos em uma rede de Petri ordinária, apresentado em Aalst [AAL98], representando um sistema para processar reclamações. A modelagem inicia com o registro de uma reclamação; após o registro são executadas paralelamente: o envio de um questionário para o reclamante e a análise da reclamação. O questionário é processado se o reclamante retorná-lo dentro de um prazo de 2 semanas; caso contrário o questionário é descartado. Com base no resultado da avaliação, a reclamação pode ou não ser processada. Esse processamento somente acontece depois da ocorrência do processamento do questionário ou da expiração do prazo de 2 semanas (*time_out*). Após, é realizada uma verificação e, caso o processamento esteja correto, a reclamação é arquivada. A existência de problemas faz com que o processamento seja reiniciado.

Uma transição pode estar associada a um ou mais lugares de entrada e a um ou mais lugares de saída. Por exemplo, na rede da Figura 3, o conjunto dos lugares de entrada da transição "*send_questionnaire*" é {c1} e o conjunto dos lugares de saída da transição "*register*" é {c1, c2}.

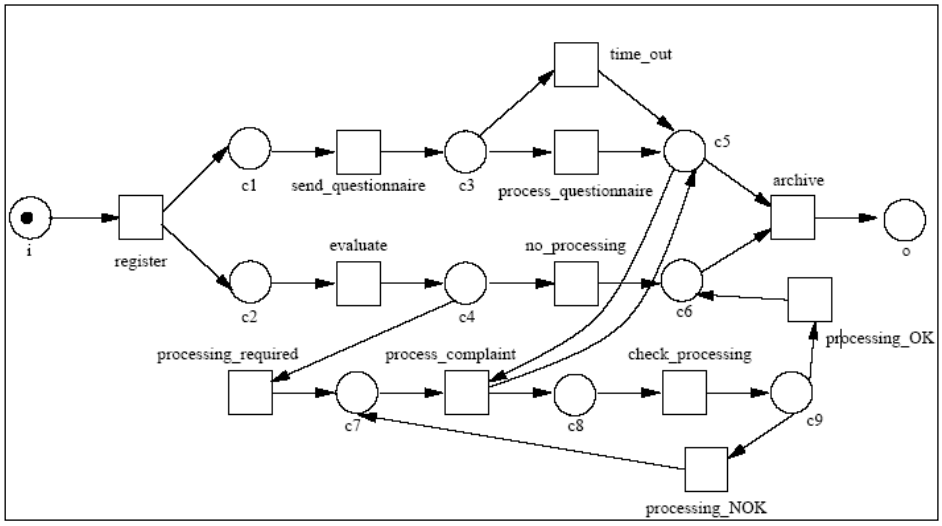

Figura 3 – Rede de Petri para o processamento de reclamações [AAL98]

A quantidade de marcas e sua distribuição nos lugares controlam o disparo das transições. Uma transição está habilitada se o lugar possui um número de marcas superior ou igual ao peso do arco que liga o lugar à transição.

Quando uma transição é executada cada marca é removida do seu lugar de entrada (de acordo com o peso dos arcos que ligam os lugares de entrada à transição) e cada uma delas é adicionada em cada um dos seus lugares de saída (de acordo com o peso dos arcos que ligam a transição aos seus lugares de saída). A marcação de uma rede de Petri muda a cada disparo de transição, permitindo simular o comportamento dinâmico do sistema e definir o seu estado em um dado momento [AAL98]. Um aspecto importante na execução de uma transição é que não há dependência quantitativa entre as marcas de antes e depois da sua execução.

2.3.2 Redes de Petri Coloridas

Jensen, em [JEN97], apresenta uma breve introdução sobre redes de Petri coloridas (CP-net), uma das extensões de redes de Petri, a qual é utilizada nesta Tese para representar o padrão proposto. O exemplo da Figura 4, modelado pelos autores com CPN Tools [CPN10], descreve um simples protocolo de transporte, transferindo pacotes em uma rede confiável (*Network*) de um remetente (*Sender*) para um receptor (*Receiver*). As elipses e círculos são chamadas de *lugares*. Eles descrevem os estados do sistema. Os retângulos são chamados de *transições*. Eles descrevem as ações. As setas são chamadas de *arcos*. As *expressões de arco* descrevem como o estado de uma CP-net é alterado quando uma transição ocorre.

Cada lugar contém um conjunto de marcações chamadas *tokens*. Em contraste com redes de Petri de baixo nível, cada um desses *tokens* carrega um valor de dado, que é de um determinado *tipo*. Como exemplo, o lugar *Send* (no lado esquerdo superior na Figura 4) tem sete *tokens* no seu estado inicial. Todos os valores dos *tokens* são do tipo INTxDATA e representam sete pacotes que serão lidos para serem enviados. O primeiro elemento é o número do pacote, enquanto o segundo é o conteúdo do pacote. Todos os 1's indicam que existe exatamente um de cada um dos pacotes para envio, pois em geral um lugar pode ter muitos *tokens* com o mesmo valor. Cada um dos lugares *NextSend* e *NextRec* iniciam com um *token* simples com valor 1, do tipo INT. Esses lugares representam dois contadores, guardando o número do próximo pacote a ser enviado/recebido. O lugar *Received* inicia com um *token* que contém uma string vazia " ", do tipo DATA. Representa o conteúdo das mensagens que tenham sido recebidas com sucesso. Os quatro lugares restantes: A, B, C e D não têm *tokens* no estado inicial. Eles representam buffers de entradas e saída da rede de transmissão.

Redes de Petri Coloridas tem esse nome porque permitem o uso de *tokens* que carregam valores de dados e podem se distinguir um dos outros, diferente dos *tokens* de Redes de Petri de baixo nível, que são desenhados como um ponto preto. No início, somente conjuntos de cores

pequenos e não estruturados eram utilizados, enumerando um conjunto fixo de processos, por exemplo. Mais tarde, percebeu-se que era possível generalizar a teoria e as ferramentas, de tal forma que os tipos de dados complexos pudessem ser utilizados como conjuntos de cores. Atualmente é comum se ter *tokens* que carregam um valor de dado complexo, por exemplo, uma lista de muitos milhares de registros, envolvendo campos de diferentes tipos. Já não existe qualquer diferença entre um conjunto de cores e um tipo, e não há diferença entre a cor de um *token* e o valor de um *token*. [JEN97]

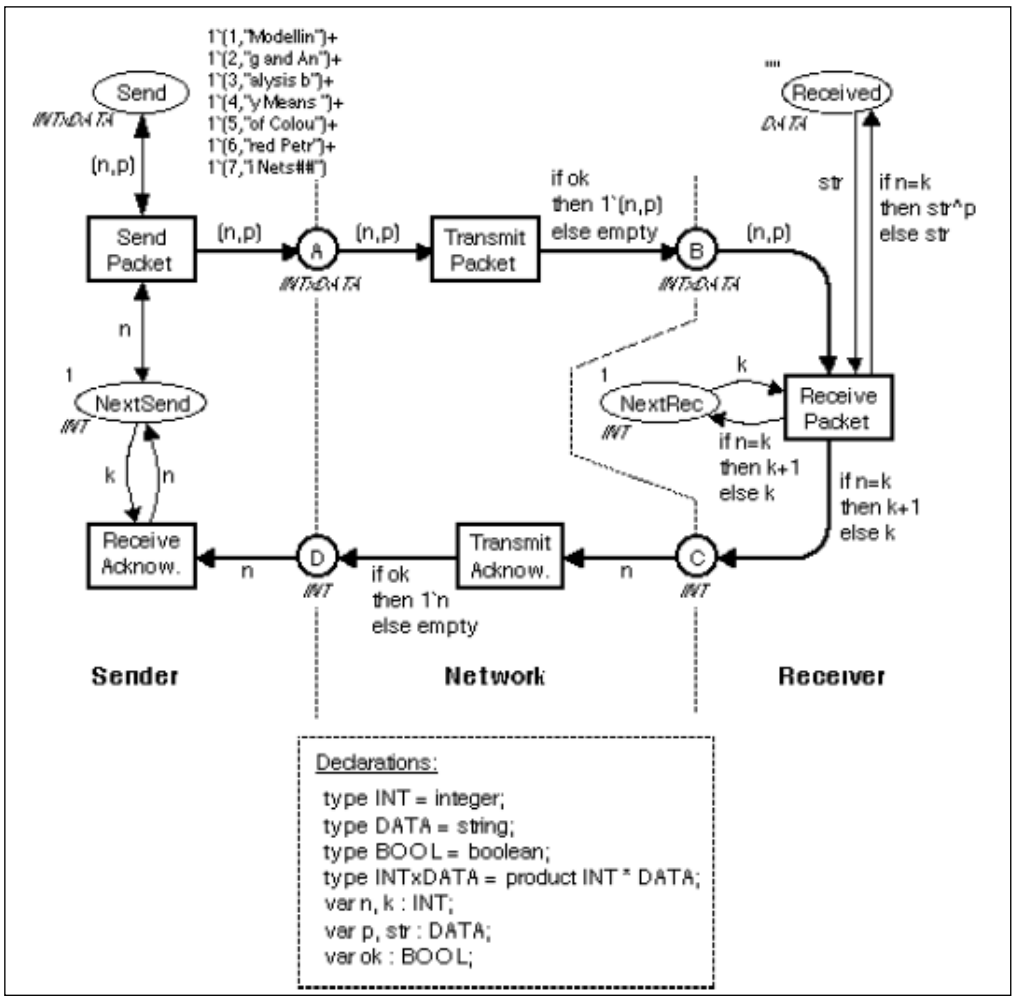

Figura 4 – Exemplo de Rede de Petri Colorida [JEN97]

Para estar habilitada para ocorrer, uma transição deve ter *tokens* suficientes nos seus lugares de entrada e esses *tokens* devem ter valores que correspondam às expressões dos arcos. Como exemplo, considera-se a transição *SendPacket* na Figura 4. Ela possui três arcos, dois dos quais são bidirecionais. As três expressões de arcos envolvem a variável n de tipo INT e a variável p de tipo DATA. Para que a transição ocorra, deve-se vincular essas duas variáveis a valores dos mesmos tipos, de tal maneira que a expressão de arco, de cada arco de entrada, é avaliada como um valor de *token* que está presente no lugar de entrada correspondente. Desde que *NextSend*

somente contém um *token* com valor 1, é óbvio que n deve possuir o valor 1. Seguindo, nota-se que p deve ser vinculado ao dado “Modellin”, pois *Send* somente tem um *token*, cujo valor do primeiro elemento do par é 1. Com a ocorrência de <n=1,p=”Modellin”> a transição *SendPacket* é habilitada, porque existe um *token* 1 no lugar *NextSend* e um *token* (1,”Modellin”) no lugar *Send*. Quando a transição ocorre, ela remove os dois *tokens* especificados dos lugares de entrada, mas logo os coloca de volta, devido aos arcos bidirecionais. Simultaneamente ela produz uma cópia do *token* (1, “Modellin”) no lugar A. Intuitivamente isso significa que o primeiro pacote foi enviado pela adição no *buffer* de entrada da rede. O pacote não é excluído de *Send* ou atualizado o contador do *NextSend,* isso porque o pacote pode ser perdido na rede, necessitando de retransmissão. O protocolo é pessimista, já que mantém o mesmo pacote para retransmissão até que receba um aviso, solicitando um novo pacote. [JEN97]

Quando o *token* (1,”Modellin”) é colocado no lugar A, a transição *TransmitPacket* é habilitada com duas ligações diferentes: <N=1, p=”Modellin”>, satisfazendo a condição, ou seja, ok igual a true e <n=1, p=”Modellin”>, sem satisfazer a condição, ou seja, ok igual a false. Se a primeira ligação for a escolhida, o pacote é transferido do lugar A para o B. Se a segunda ligação for a escolhida, o pacote é perdido na rede, representado pela expressão *empty*. A escolha entre as duas ligações é não-determinística (na simulação de interação a escolha pode ser feita pelo usuário e na simulação automatizada é feita por um gerador randômico). No exemplo há, portanto, 50% de chance de sucesso/falha. De qualquer forma é fácil modificar a CP-net para utilizar uma probabilidade diferente.

A transição *SendPacket* permanece habilitada e dela uma retransmissão pode ocorrer a qualquer momento. Os *tokens* de um determinado lugar podem ser utilizados em qualquer ordem, independente da ordem de sua chegada. Isso significa que os pacotes podem ultrapassar uns aos outros, tanto para o lugar A quanto para o lugar B.

Quando um *token* chega ao lugar B, a transição *ReceivePacket* é habilitada. Ela compara o número n no pacote de entrada ao número k no contador *NextRec*. Se os dois valores forem iguais, o pacote é o esperado. Assim, o conteúdo do dado p do pacote é adicionado ao conteúdo do *token str* no lugar *Received*, utilizando o operador de concatenação de strings ^, acrescenta-se uma unidade ao contador *NextRec* e um reconhecimento é enviado pelo lugar C, contendo o número do próximo pacote a ser recebido. Se os valores forem diferentes o pacote é ignorado. O valor do *token* nos lugares *Received* e *NextRec* são inalterados.

A transição *TransmitAcknow* trabalha de forma similar ao *TransmitPacket*. Ela transmite os reconhecimentos sobre a rede, movendo-os do lugar C para o D, a menos que *ok=false*, caso no qual o reconhecimento seria perdido. A transição *ReceiveAcknow* manipula esses reconhecimentos que estão ao alcance do remetente. Atualiza o contador *NextSend*, assim que o remetente iniciar a transmissão de um novo pacote.

Jensen, [JEN94], também apresenta a definição formal e o comportamento de redes de Petri Coloridas. Uma rede desse tipo é definida como uma tupla múltipla, sendo entendida com o propósito de oferecer uma notação matemática, sem ambiguidades, sobre a definição de redes de Petri e sua semântica. Qualquer rede concreta será sempre especificada por meio de um diagrama. Pode-se ter como princípio, mas não como prática, a facilidade de se traduzir o diagrama de uma rede de Petri Colorida dentro de uma tupla de Rede de Petri e vice-versa. A especificação dessa tupla é adequada quando se quer formular definições gerais e prover teoremas aplicados a todas as classes de redes de Petri Coloridas ou a parte delas. Já a especificação em forma de grafo é mais adequada quando se quer construir a modelagem de um sistema específico.

2.3.2.1 Definição de Redes de Petri Coloridas

Para uma definição abstrata de Redes de Petri Coloridas não é necessário se fixar apenas na sintaxe de modelagem da rede. Assume-se que a sintaxe, com semânticas bem definidas, torna possível a definição de caminhos não ambíguos. Define-se, portanto, com base em Jensen, [JEN94]:

- elementos de um tipo T: o conjunto de todos os elementos em T é representado pelo próprio nome de T;
- tipo de uma variável v é representado por Tipo(v);
- tipo de uma expressão, expr é representado por Tipo(expr);
- conjunto de variáveis em uma expressão, expr é representado por Var(expr);
- ligação de um conjunto de variáveis, V, associando com cada variável v Є V um elemento b(v) Є Tipo(v), sendo b um elemento qualquer da rede;
- valor obtido pela avaliação de uma expressão expr, em uma ligação b é representado por expr<b>. Var(expr) é requerido para ser um subconjunto de variáveis de b, e a avaliação é realizada substituindo cada variável v Є Var(expr) pelo valor b(v) Є Tipo(v) determinado pela ligação.

Uma expressão sem variáveis é chamada de expressão fechada. Ela pode ser avaliada em todas as ligações e, em todas elas, fornece o mesmo valor, o que significa a própria expressão. Portanto, algumas vezes escrever expr é o mesmo que escrever expr<b>.

Definição de Rede de Petri Colorida. Uma CP-net é uma tupla CPN = <Σ, P, T, A, N, C, G, E, I> onde:

(i) Σ é um conjunto finito de tipos não vazios também chamados conjuntos coloridos.
O conjunto de tipos determina os valores de dados e as operações e funções que podem ser usadas nas próximas expressões, ou seja, nas expressões de arcos, guardas e inicializações. Assume-se que cada tipo tem, pelo menos, um elemento.

(ii) P é um conjunto finito de lugares não vazio.

(iii) T é um conjunto finito de transições não vazio.

(iv) A é um conjunto finito de arcos nos quais: $P \cap T = P \cap A = T \cap A = \emptyset$.
Os lugares, transições e arcos são descritos pelos três conjuntos P, T e A os quais devem ser finitos e pares disjuntos. Pela requisição dos conjuntos de lugares, transições e arcos de serem finitos, evita-se a possibilidade, por exemplo, de se ter um número infinito de arcos entre dois nodos.

(v) N é uma função de nodo. Ele é definido por A dentro de $P \times T \cup T \times P$.
A função de nodo mapeia cada arco dentro de um par onde o primeiro elemento é o nodo origem e o segundo elemento é o nodo destino. Os dois nodos devem ser de tipos diferentes, isto é, um deve ser um lugar, enquanto o outro é uma transição. Assume-se que uma rede de Petri Colorida pode ter vários arcos entre um mesmo par de nodos ordenados.

(vi) C é uma função colorida. Ela é definida por P dentro de Σ.
A função colorida C mapeia cada lugar p para um tipo C(p). Intuitivamente, isto significa que cada *token* em p deve ter um valor de dado que pertença a C(p).

(vii) G é uma função de guarda. Ela é definida por um T dentro de expressões tais que: $\forall t \in T$: [Tipo(G(t)) = B ^ Tipo(Var(G(t))) $\subseteq$ Σ], onde B representa um tipo booleano.
A função de guarda G mapeia cada transição t dentro de uma expressão booleana onde todas as variáveis têm tipos que pertencem a Σ. Quando se cria esquemas em redes de Petri coloridas omitem-se as expressões de guarda, assumindo-se que sejam verdadeiras.

(viii) E é uma função de expressão de arco. Ela é definida por um A dentro de expressões tais que: $\forall a \in A$: [Tipo(E(a)) = $C(p)_{MS}$ ^ Tipo(Var(E(a))) $\subseteq \Sigma$] onde p é um lugar de N(a) e MS são todos os múltiplos conjuntos de C.

A função de expressão de arco E mapeia cada arco **a** dentro de uma expressão de tipo $C(p)_{MS}$. Isso significa que cada expressão de arco deve avaliar os múltiplos conjuntos sobre o tipo de lugar adjacente, p. Permite-se que um diagrama de rede de Petri colorida tenha uma expressão de arco *expr* de tipo C(p), e considera-se que esse seja um atalho para 1'(expr).

(ix) I é uma função de inicialização. Ela é definida por P dentro de expressões fechadas tais que: $\forall p \in P$: [Tipo(I(p)) = $C(p)_{MS}$].

A função de inicialização I mapeia cada lugar p dentro de uma expressão a qual deve ser de tipo $C(p)_{MS}$. Quando se desenha uma rede de Petri colorida omitem-se as expressões de inicialização que seriam avaliadas como ∅.

2.3.2.2 Aplicação de Redes de Petri Coloridas

Nardi [NAR09] afirma que as redes de Petri são intuitivas e de fácil compreensão para sistemas pequenos. Identifica, como exemplo, um processo de fabricação de dois modelos de veículos utilizando três tipos de recursos, conforme demonstrado na Figura 5. O autor afirma que "extrapolando este exemplo para a produção de trinta modelos usando quinze tipos de recursos, chega-se a uma rede muito menos legível, dado o número de repetições de atividades e estados. Nesse caso, enquanto há duas atividades T1 (T1a e T1b) na rede da Figura 5**,** para o novo cenário haveria até trinta atividades T1, sendo uma para cada modelo, tornando-se praticamente ilegível para sistemas grandes ou complexos". As redes de Petri coloridas surgiram com o objetivo de suprir essa dificuldade.

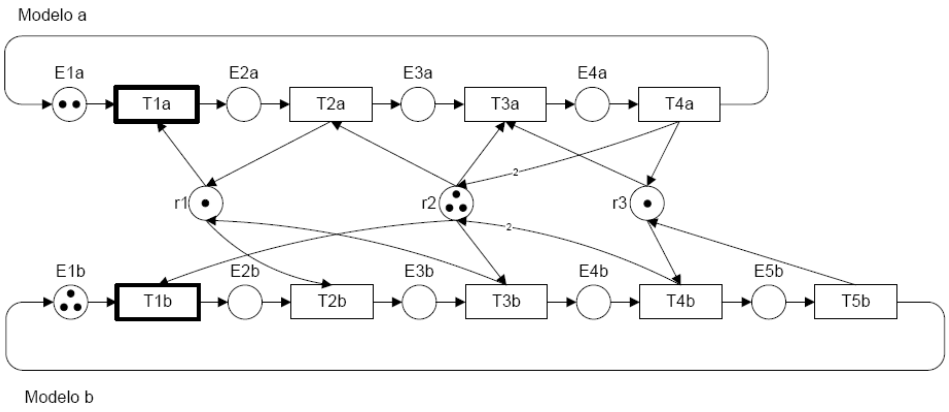

Figura 5 – Rede de Petri Exemplo – Fabricação de dois Modelos de Veículos [NAR09]

Conforme Girault et al. [GIR02] e Penha et al. [PEN04] o principal objetivo das Redes de Petri Coloridas é “a redução do tamanho do modelo, permitindo que *tokens* individualizados (coloridos) representem diferentes processos ou recursos em uma mesma sub-rede”. Na definição inicial do padrão, os *tokens* eram representados por cores ou mesmo por padrões que possibilitassem a distinção dos *tokens*. Visando aumentar a abrangência, os *tokens* são representados por estruturas de dados complexas, contendo informações diferenciadas por tipos, possibilitando que os arcos realizem operações sobre eles.

A Figura 6 ilustra uma rede de Petri para um sistema de manufatura. Esse mesmo sistema é apresentado em rede de Petri Colorida na Figura 7. A rede apresentada na Figura 6 modela um sistema de manufatura com dois processos, compartilhando duas máquinas. A adição de um novo processo, por exemplo, mesmo compartilhando as mesmas máquinas, exige que sejam modelados todos os estágios desse novo processo, por meio de estados, ações e relações entre este processo e os dois já existentes, e entre este processo e os recursos disponíveis. Isso faz com que a rede de Petri que modela o sistema cresça significativamente, mostrando a conveniência em usar redes de Petri Coloridas, pois independente do número de processos, a estrutura continua a mesma, conforme visualizado na Figura 7 [PEN04].

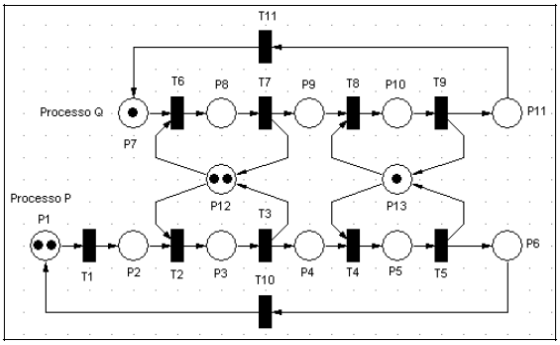

Figura 6 – Rede de Petri Exemplo – Sistema de Manufatura [PEN04]

Na modelagem de *workflow*, utilizando redes de Petri, cada tarefa é representada por uma transição correspondente. Lugares representam as pré e pós-condições ou, ainda, os recursos requeridos para executar determinada tarefa. Os arcos representam relações lógicas entre as tarefas e o próprio fluxo de trabalho. A representação gráfica das redes de Petri tem se mostrado muito útil, pois permite a visualização dos processos e a comunicação entre eles [AAL98].

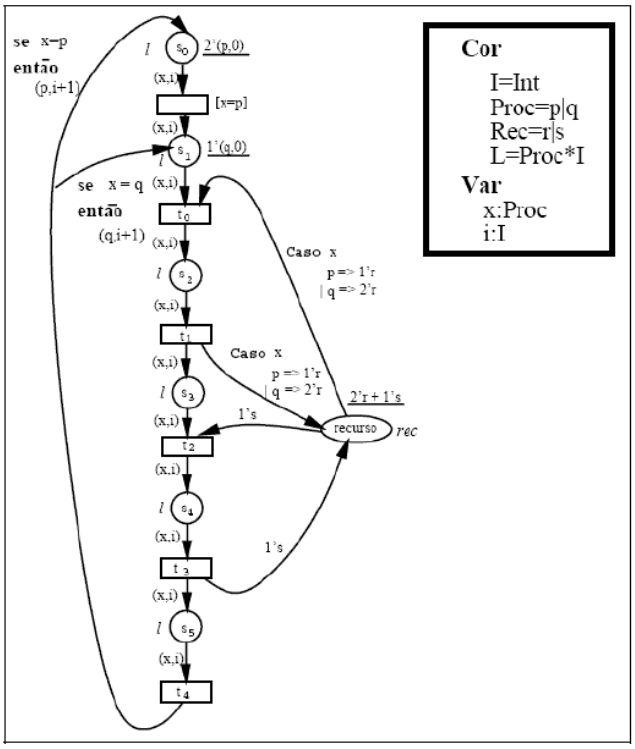

Figura 7 – Rede de Petri Colorida Exemplo – Sistema de Manufatura [MAC96] Apud [PEN04]

Em Aalst et al. [AAL98, AAL02] são encontradas três razões principais para aplicar redes de Petri na modelagem de *workflow*:

- as redes de Petri possuem tanto semântica formal quanto natureza gráfica;
- elas podem modelar explicitamente estados do sistema; e
- há variedade e disponibilidade de técnicas de análise.

2.4 Considerações do Capítulo

Para Gil et al. [GIL07], um tema importante de investigação está na possibilidade de *workflows* científicos serem construídos sobre tecnologias já existentes para a criação de *workflows* de negócios ou da exigência de existência de novas abordagens para essa criação. Os autores afirmam que *workflows* científicos e de negócios não são facilmente distinguíveis, pois existem exemplos que apresentam características em comum: grande volume de dados, grande quantidade de paralelismo etc. Por outro lado, os *workflows* científicos requerem modelagem flexível e capacidade exploratória que os distinguem dos *workflows* de negócios. Outra característica de *workflows* científicos é a variedade e heterogeneidade de dados executados sobre um simples *workflow*. Os *workflows* científicos, conforme os autores, apóiam discursos científicos detalhados tão bem quanto possibilitam a execução de processos repetitivos. Além disso, deve-se prover uma visão estável do sistema, mesmo com as constantes mudanças de tecnologia e plataforma. Isso quer dizer que uma vez obtidos resultados, esses devem ser os

mesmos obtidos com uma nova execução anos mais tarde, se necessário (proveniência). O desafio de se utilizar *workflows* em aplicações científicas é ainda maior pela necessidade de se validar os resultados gerados pelos dados e compartilhar essas informações com a comunidade científica. Assim, a rastreabilidade e o compartilhamento são requisitos fundamentais de *workflows* científicos.

Para Davidson e Freire [DAV08] sistemas de *workflow* auxiliam cientistas a conceituarem e gerenciarem processos, permitindo a criação e reuso de tarefas de análise, permitindo também o processo de descoberta pelo gerenciamento de dados utilizados e gerados em cada etapa e sistematizando a informação para uso posterior. Sistemas de *workflows* científicos utilizam-se, tipicamente, de modelos computacionais simples, modelos de fluxos de dados, onde a ordem de execução é determinada pelos fluxos de dados do *workflow*. Esse é um contraste em relação a *workflows* de negócios, os quais necessitam de expressivas linguagens de definição para especificar complexos fluxos de controle. Os sistemas de *workflow* apresentam um grande número de vantagens para execução de aplicações científicas: conferem um modelo de programação simples, baseado na sequenciação de tarefas, compostas pela conexão de entradas e saídas e interfaces gráficas para criação visual dos fluxos.

Machado [MAC07a] apresenta algumas diferenças entre *workflows* de negócios e *workflows* científicos discutidas por Meyer [MEY04] e Weske [WES95]. Esses autores confirmam algumas das diferenças já descritas e apresentam outras:

- *workflows* de negócio são modelados para atender a um processo relativamente fixo, porém no caso de *workflows* científicos a definição do *workflow* envolve a tomada de diversas decisões, análises e, muitas vezes, trabalho em equipe;
- enquanto *workflows* de negócio são orientados pelo fluxo de controle das atividades, *workflows* científicos são orientados pelo fluxo de dados;
- *workflows* de negócio requerem poucas mensagens de coordenação e troca de documentos e dados entre as atividades, enquanto que *workflows* científicos utilizam muitos dados, geralmente derivados de diferentes fontes e nenhum documento é modificado;
- para *workflows* científicos tanto as respostas positivas quanto as negativas precisam ser analisadas e por isso devem ser armazenadas;

- em *workflows* de negócio o modelo desenvolvido não é alterado durante a execução do *workflow* devido aos resultados obtidos após a execução de cada etapa. Já a definição de *workflows* científicos é um processo dinâmico, influenciado muitas vezes por resultados obtidos durante a execução, gerando constantes mudanças no fluxo de execução.

A Figura 8 apresenta uma tabela comparativa, distinguindo *workflows* de negócios dos científicos. Com base em Mattoso et al. [MAT08], a tabela foi adaptada com foco em características da área de aplicação, apresentada no Capítulo 3, enfatizando características direcionadas a *workflows* científicos.

Caraterísticas	Científico	Comercial
Desenho dos workflows	Cientistas e pesquisadores	Especialistas no negócio ou profissionais de informática
Conexões entre as tarefas	Dados (maior parte), Controle (menor parte)	Dados (menor parte), Controle (maior parte)
Tipos de dados	Arquivos heterogêneos e altamente complexos	Documentos bem estruturados, em geral pequenos
Volume de dados manipulados	Alto	Baixo
Freqüência de mudanças nos Wfs	Elevada	Baixa
Validação de dados intermediários	Sim, checagem de consistência e validações para a continuidade do processamento	
Aderência a padrões de mercado	Baixa	Moderada
Execução parcial	Desejável e necessário	
Modificação dinâmica	Desejável. Alterações podem ser realizadas em função de resultados obtidos	
Suporte ao reuso	Desejável, parte de um Wf pode ser aproveitado por outros pesquisadores no mesmo experimento ou em experimentos futuros	Não é crucial, os Wfs são criados para tarefas específicas
Suporte a proveniência de dados	Sim, É desejável que se armazenem informações sobre as fontes de dados utilizadas, sobre as execuções dos Wfs e seus resultados produzidos	
Aplica-se a um domínio específico	Em alguns casos sim, porém existem muitos exemplo de workflows interdisciplinares	Sim, sempre

Figura 8 – Comparativo entre *workflows* de negócios e *workflows* científicos (Adaptada de [MAT08])

Este capítulo, além de apresentar a diferenciação entre *workflows* científicos dos de negócios, possibilitando o entendimento e a forma de aplicação para a realidade estudada, apresentada no capítulo 3, também apresentou conceitos de um dos formalismos utilizados para representação e modelagem de *workflows*, independente do tipo. A utilização de formalismos para a representação de situações da realidade é frequente. Aalst [AAL01] afirma que o uso de um

conceito formal possui inúmeras vantagens, principalmente tratando-se de representação de processos:

- exigência de ter uma definição precisa;
- ambiguidades, incertezas e contradições são evitadas, ao contrário de muitas técnicas de diagramação informal;
- o formalismo pode ser usado para argumentar sobre o processo.

Este capítulo também apresentou redes de Petri e uma de suas extensões, as redes de Petri Coloridas. Sobre esses formalismos, Kotb e Baoumgart, em [KOT05], afirmam que redes de Petri podem ser usadas para especificar o roteamento de instâncias de *workflow* e que uma rede de Petri pode ser usada, também, para modelar um problema, representando-o por meio da estrutura de rede. Gubala e Bubak, em [GUB06], utilizam as redes de Petri coloridas para modelar *workflows* científicos e afirmam que existem algumas vantagens em utilizá-las. Entre elas estão as seguintes:

- a possibilidade de se utilizar diferentes construtores, com grande poder de expressão;
- a existência de algoritmos documentados e ferramentas de modelagem;
- a facilidade de aplicação em diferentes domínios, como em *workflows* científicos;
- a noção de *tokens* faz com que se tenha controle e manutenção de estados da aplicação durante a execução.

Nesse contexto, o estudo de *workflows* e de um formalismo utilizado para sua representação é fundamental para o entendimento do padrão apresentado nesta Tese no Capítulo 5. As redes de Petri coloridas, em especial, são utilizadas para especificar o padrão desenvolvido e seu funcionamento. No Capítulo 3, algumas características da área de aplicação, a Bioinformática, são apresentadas, bem como estudos realizados pelo LABIO (Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas), norteadores deste trabalho.

3 ÁREA DE APLICAÇÃO

Este capítulo apresenta a área de Bioinformática, algumas de suas características e desafios. Também são apresentadas, neste capítulo, atividades desenvolvidas no LABIO (Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas) da Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), com ênfase em algumas linhas de estudo: flexibilidade de macromoléculas, definição de *workflow* científico para a automatização de processos e classificação dos dados.

3.1 Bioinformática

A área com o qual o padrão apresentado nesta Tese é validado é a área de Bioinformática. Para Mattoso et al. [MAT08], a Bioinformática é uma área de conhecimento convergente e multidisciplinar que envolve o intenso uso de ferramentas computacionais. É uma área na qual se busca resolver problemas do cotidiano, principalmente os que envolvam manipulação de grandes bancos de dados de genomas, sequências protéicas, entre outras. Os autores também acreditam que essa área possui como objetivo final a coleta, a organização, o armazenamento, a recuperação e as análises de dados biológicos, propiciando a inferência ou descoberta de informações sobre a biologia e sobre a evolução dos organismos.

Weske et al., em [WES95], definem que os experimentos científicos, independentemente do domínio de aplicação, devem obedecer a um conjunto rígido de requisitos. Esses requisitos, válidos até os dias atuais são:

- Deve ser possível reproduzir e disseminar experimentos: os resultados científicos devem ser reproduzidos e seus resultados amplamente disseminados, de forma que diferentes cientistas possam comprová-los. Para que isso aconteça, os experimentos devem ser bem documentados.
- Experimentos devem ser projetados por meio da utilização de um protocolo, ou metodologia, permitindo que possam ser conduzidos a partir de um ponto inicial

ou por meio da combinação e reutilização de experimentos já realizados anteriormente.

- Experimentos devem ter sua condução controlada, uma vez que cada experimento consiste em um número de etapas com restrições complexas, podendo ser executadas em paralelo, necessitando de maior controle caso exista a necessidade de se refazer uma determinada etapa do processo.

Coutinho et al., em [COU10], afirmam que, por muitos anos, cientistas da área de Bioinformática têm manipulado um grande volume de dados e que seus estudos baseiam-se, principalmente, em experimentos por meio de simulação computacional, ou seja, experimentos *in silico,* e que demandam grande capacidade de processamento por parte dos computadores [TRA03]. Mattoso et al. [MAT08] afirmam que os experimentos *in silico* fazem intensos usos de várias ferramentas computacionais, podendo ser utilizadas de forma encadeada, contemplando todas as etapas necessárias para a finalização dos experimentos. Esse uso intenso de várias ferramentas computacionais, quando utilizadas de forma encadeada, dá origem aos *workflows* científicos, já apresentados no capítulo anterior.

Para Machado [MAC07] o avanço da biologia molecular e das ferramentas de simulação *in silico*, fez com que o planejamento de medicamentos fosse realizado de maneira mais lógica, sendo chamado de Desenho Racional de Fármacos [KUN92]. A interação entre moléculas é o princípio fundamental do planejamento racional de fármacos. Para tanto, tem-se receptores e ligantes. Machado [MAC07a] trata proteínas, macromoléculas e receptores como sinônimos e afirma que, assim como enzimas e DNA não são rígidas em seu ambiente celular. Essa flexibilidade deve ser explicitamente considerada durante o processo de desenvolvimento de novos fármacos (*drug design*). Além disso, afirma que uma das principais etapas desse processo de desenvolvimento de novos fármacos é a docagem molecular, na qual se investiga e avalia o melhor encaixe do ligante na estrutura alvo ou receptor. As ligações ocorrem em locais específicos: sítios de ligação, cavidades ou regiões na superfície das moléculas onde existe um ambiente favorável à interação ligante-receptor. A partir de ligantes identificados como promissores nos experimentos *in silico*, experimentos *in vitro* podem ser realizados com, talvez, maior possibilidade de êxito.

Um dos grandes desafios está, justamente, em manipular essa grande quantidade de dados [LUS01]. Bancos de dados de pequenas moléculas (ligantes), como o ZINC [IRW05], têm disponíveis mais de 20 milhões de compostos [ZIN09]. A análise, mesmo que *in silico*, da interação

de todos esses possíveis compostos com uma determinada proteína-alvo (receptor) e sua respectiva conformação, se torna inviável de ser executada, pois se estima que seriam necessários aproximadamente 62 trilhões de minutos (aproximadamente 117 mil anos) até o término da execução de todos os experimentos, onde para cada experimento receptor-ligante, chamado de docagem molecular (*docking*), considera-se um tempo mínimo por execução de 1 minuto, o que nem sempre é verificado, pois o tempo tende a ser maior, dependendo do equipamento a ser utilizado.

Buscando melhorar esse tempo e otimizar o processo de docagem molecular, o LABIO, Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas, reúne pesquisadores que investigam e pesquisam novas soluções.

3.2 LABIO: Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas

O Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas – LABIO – da Faculdade de Informática da PUCRS tem se dedicado a estudar e realizar experimentos *in silico* sobre a enzima InhA do *Mycobacterium tuberculosis* (MTB). Esse estudo deve-se pelo crescimento mundial da tuberculose [WOR06a, WOR06b] e, também, pelo aumento dos casos de tuberculose resistentes à isoniazida, um dos principais fármacos utilizados no seu tratamento. O objetivo do LABIO é contribuir para o desenvolvimento de novos fármacos para o tratamento da tuberculose, por meio do aumento de qualidade da seleção de compostos candidatos. Esforços da indústria farmacêutica, combinados com avanços na área de genômica estrutural, permitiram que um grande número de estruturas de proteínas e pequenas moléculas estejam disponíveis em repositórios [SIN06]. Essas estruturas são de fundamental importância para o entendimento dos sistemas biológicos, assim como para o desenvolvimento mais eficiente de novos fármacos.

3.2.1 Flexibilidade de Macromoléculas

Como as macromoléculas não são rígidas em seu ambiente celular, tona-se muito interessante que sua flexibilidade seja levada em consideração em um processo de docagem molecular [MAC07]. Um dos trabalhos desenvolvidos dentro do LABIO, por Machado em [MAC07a], apresenta que experimentos de docagem molecular podem ser executados por diferentes softwares, como por exemplo: o AutoDock3.05 [GOO96], o FLEXX [RAR96] e o DOCK4.0 [EWI01]. A autora também afirma que a maioria desses softwares trata a flexibilidade do ligante, mas apresenta dificuldades em considerar a flexibilidade do receptor. Os softwares que consideram a flexibilidade do receptor fazem isso somente de uma maneira muito limitada. A autora, com o objetivo de contornar esse problema e incluir uma representação mais realista da

flexibilidade natural dos receptores durante os experimentos de docagem, considerou um conjunto de *snapshots* do receptor gerados por uma simulação da dinâmica molecular [LIN02] onde, para cada possível conformação do receptor (*snapshot*), um experimento de docagem deve ser executado e analisado. Uma conformação de um receptor é um conjunto de coordenadas X, Y, Z para cada átomo do mesmo, em Angstrons (Å), além de informações topológicas referentes a cada um deles. Para a enzima InhA que tem 4008 átomos, em cada *snapshot* tem-se 12.024 valores reais. Assim, nos experimentos de docagem molecular, são testadas as diversas conformações do ligante com as diversas conformações do receptor. Desta maneira, podem ser executados vários experimentos utilizando, em cada um, uma das conformações (*snapshot*) de determinado receptor e um mesmo ligante a se ligar a esse receptor. Os testes iniciais apresentavam tempo aproximado de 5 minutos, considerando apenas uma conformação do receptor para um ligante. Por exemplo, se uma simulação de 3 ns (3*10-9 segundos) capturar a conformação do receptor a cada 0,5 ps, ter-se-ia 6.000 conformações para testar com um único par receptor-ligante. A Figura 9 ilustra o processo de docagem molecular.

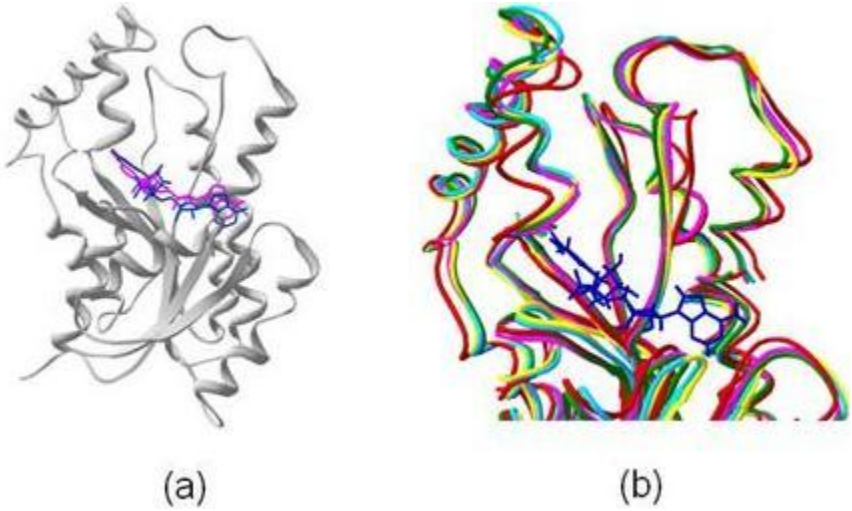

Figura 9 – (a) Representação esquemática do processo de docagem molecular em 3D. A proteína é representada na forma de ribbons, em cinza, e o ligante em linhas (magenta e ciano). (b) Flexibilidade do sistema InhA-NADH em diferentes momentos ao longo de uma simulação por dinâmica molecular. Sobreposição de diferentes conformações da InhA (ciano, amarelo, magenta e verde) gerado por dinâmica molecular em [SCH05]. Figura extraída de [MAC07]

Até o momento, a identificação de ligantes promissores se fez somente pela energia livre de ligação (*Free Energy of Binding* – FEB) obtida na docagem molecular: quanto maior a energia liberada, maior a chance do ligante ser promissor. Cabe ressaltar que experimentos com diferentes *snapshots* para o receptor podem resultar em FEBs bastante distintos e que o volume

de dados resultante de cada experimento de docagem molecular torna tal tarefa humanamente impossível sem um ferramental computacional que assuma grande parte das análises.

3.2.2 Um *Workflow* Científico para o Processo de Desenvolvimento de Fármacos

Buscando melhorar o suporte computacional para a realização desse tipo de experimento, Machado [MAC07a] desenvolveu um *workflow* científico para automatizar o processo de execução de experimentos de docagem molecular, considerando a flexibilidade do receptor. A Figura 10 e a Figura 11 ilustram a correspondente modelagem. Com isso, os experimentos puderam ser executados de forma mais automática. Antes, execuções desse tipo no LABIO eram manuais ou com o auxílio de scripts básicos, que necessitavam serem modificados a cada execução de diferentes experimentos receptor-ligante.

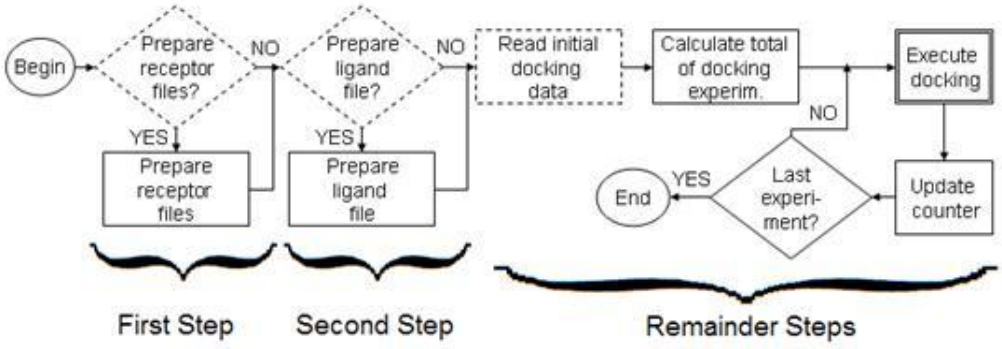

Figura 10 – Processo de CADD com flexibilidade explícita do receptor. Extraído de [MAC07]

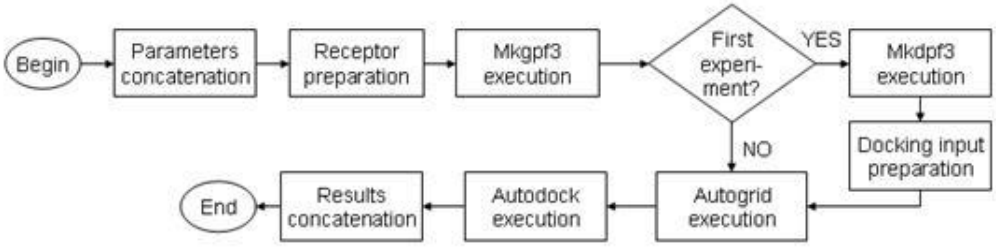

Figura 11 – Subprocesso “Execute docking”, que executa os experimentos de docagem molecular. Extraído de [MAC07]

Machado et al. reportam em [MAC07] a execução de 3 estudos de caso utilizando as conformações do receptor InhA gerados no trabalho de Schroeder et al. [SCH05] e os ligantes NADH [DES95], PIF [OLI04] e TCL [KUO03]. Cada estudo de caso utilizou um conjunto de 3.100 conformações do receptor InhA. Como, em média, cada experimento de docagem molecular despendia de 5 a 10 minutos de execução, e como em cada estudo de caso foram executados 3.100 experimentos de docagem molecular, o tempo total de cada estudo de caso passou a ser, em média, 350 horas (o tempo varia de acordo com a máquina onde a execução está sendo feita e o tamanho do receptor e do ligante). A execução automática de todo esse processo já fez com que os experimentos pudessem ser executados sem a necessidade de intervenção humana. Contudo, o

LABIO demanda por simulações que cubram entre 10 ns e 100 ns, bem mais que os 3 ns da dinâmica molecular simulada e experimentada. Também demanda por experimentos com mais ligantes, que apresentem alguma similaridade com os fármacos conhecidos.

Mantido o atual protocolo experimental de se ter uma conformação da InhA para cada picossegundo, 10.000 a 100.000 experimentos de docagem molecular seriam necessários para cada ligante a ser testado, o que demandaria muito tempo. Busca-se, assim, diminuir o número de experimentos de docagem, procurando manter a qualidade dos resultados obtidos.

3.2.3 Classificação dos Dados

A representação das conformações das moléculas tem sido feita por arquivos produzidos por simulações de dinâmica molecular, como o programa AMBER 6.0 [PEA95]. Tipicamente, esses arquivos podem ser dos formatos .PDB (Protein Data Bank) ou uma combinação dos arquivos .CRD (de coordenadas XYZ dos átomos) e .TOP (de topologia da molécula). A Figura 12 mostra trechos desses arquivos. A Figura 12(a) mostra um trecho de um arquivo .CRD, composto por um conjunto de coordenadas em sequência. A Figura 12(b) mostra o trecho inicial do arquivo .TOP correspondente ao .CRD da Figura 12(a), contendo a listagem com os nomes de cada átomo, de cada resíduo, o total de átomos da proteína e outras características. A Figura 12(c) mostra um arquivo .PDB correspondente aos arquivos .CRD e .TOP, onde os átomos encontram-se associados às suas coordenadas e informações adicionais. O formato .PDB é largamente empregado para armazenamento de conformações de moléculas em bancos de dados [BER00] e na visualização de moléculas por computador.

```
INHA + NADH de M. tubewrculosis [1ENY,27-JAN-1995] Residues 1-268
  64.433  26.825 128.851  65.342  26.818 129.291  64.348  27.752 128.459  63.721
  26.799 129.567  64.230  25.744 127.845  65.119  25.789 127.216  64.261  24.354
 128.504  63.380  24.273 129.141  64.290  23.475 127.859  65.158  24.297 129.120
  62.919  25.939 127.160  61.875  26.090 127.794  62.964  25.857 125.847  63.851
```

(a)

```
INHA + NADH de M. tubewrculosis [1ENY,27-JAN-1995] Residues 1-268
 31481    23 29437  2080  4635  2833  7940  3718     0     0 59005  9407
  2080  2833  3718    79   196   143    39     1     0     0     0     0
     0     0     0     1    71     0
N   H1  H2  H3  CA  HA  CB  HB1 HB2 HB3 C   O   N   H   CA  HA2 HA3 C   O   N
H   CA  HA  CB  HB2 HB3 CG  HG  CD1 HD11HD12HD13CD2 HD21HD22HD23C   O   N   H
```

(b)

```
ATOM      1  N   ALA     1      64.433  26.825 128.851  0.00  0.00
ATOM      2  H1  ALA     1      65.342  26.818 129.291  0.00  0.00
ATOM      3  H2  ALA     1      64.348  27.752 128.459  0.00  0.00
ATOM      4  H3  ALA     1      63.721  26.799 129.567  0.00  0.00
ATOM      5  CA  ALA     1      64.230  25.744 127.845  0.00  0.00
```

(c)

Figura 12 – (a) Trecho de um arquivo .CRD de saída da simulação por dinâmica molecular. (b) Trecho do arquivo de topologia .TOP utilizado. (c) Exemplo de arquivo .PDB. Extraído de [MAC07a].

Winck et al. [WIN09], [WIN10], modelaram um banco de dados, chamado FReDD (*Flexible Receptor Docking Database*), para dar suporte completo aos experimentos de docagem molecular e, também, para mineração de dados. Seu modelo final é composto por 16 tabelas, conforme mostrado na Figura 13. Inicialmente, FReDD amazenou dados da InhA e de 4 ligantes: NADH [DES95], TCL [KUO03], PIF [OLI04] e ETH [BAN94]. Tem-se um total de 12.424.800 coordenadas dos átomos do receptor, que correspondem aos 4,008 átomos de cada conformação (*snapshot*), multiplicado por 3.100 conformações. Além disso, tem 3.248.330 registros de coordenadas dos átomos dos ligantes, resultantes dos experimentos de docagem. A Tabela 1 sumariza essas populações. É importante destacar que a coluna “Total de Conformações”, da Tabela 1, representa que são realizadas 10 execuções de experimentos para cada par ligante.

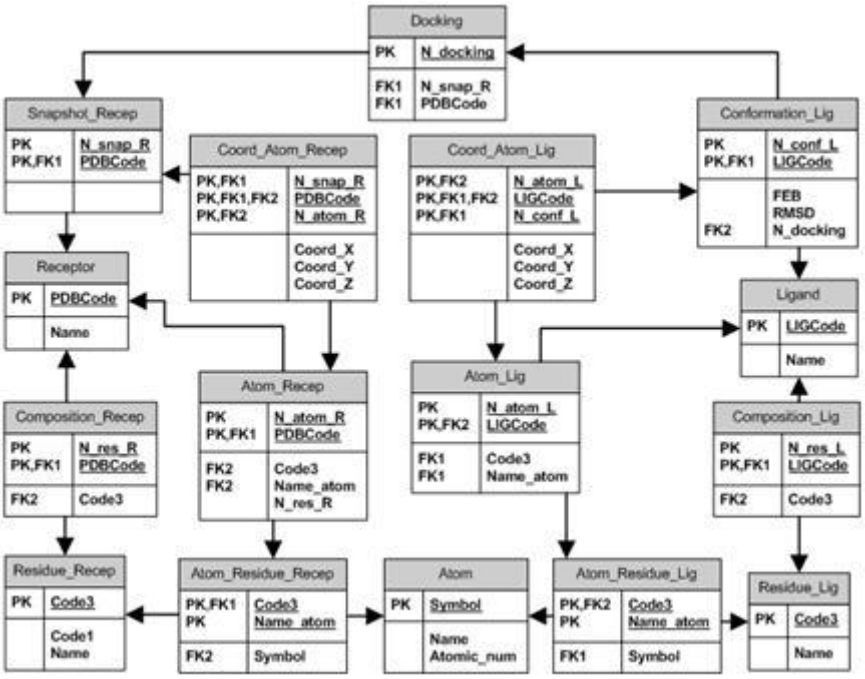

Figura 13 – Modelo final do FreDD, diagramado com Microsoft Visio. Figura extraída de [WIN09].

Na sequência, Winck et al. [WIN09] realizaram um experimento de mineração de dados, usando como atributos preditivos as distâncias mínimas entre átomos do ligante e átomos dos resíduos de aminoácidos do receptor, em cada experimento de docagem, e o valor de FEB (*free energy of binding*) como atributo alvo, a ser predito. O trabalho desenvolvido pelos autores pode ser utilizado por Karina Machado para a determinação de uma função de similaridade, assunto de sua Tese de Doutorado, a qual pode ser aplicada ao padrão definido nesta Tese.

Tabela 1 – Conformações geradas por dinâmica molecular e resultados de docagem com NADH [DES95], TCL [KUO03], PIF [OLI04] e ETH [BAN94], armazenados inicialmente no FreDD [WIN09].

Receptores e Ligantes	Número de átomos	Número de docagens válidas	Total de conformações	Total de Coordenadas
NADH	52	3.100	31.000	1.612.000
TCL	18	2.837	28.370	510.660
PIF	24	3.042	30.420	730.080
ETH	13	3.043	30.430	395.590
Total		12.022	120.220	3.248.330

3.3 Considerações do Capítulo

Mattoso et al. em [MAT09] afirmam que a ciência tem feito cada vez mais uso de procedimentos computacionais, buscando lidar com o aumento constante dos volumes de dados e manipulações necessárias aos experimentos científicos. Também afirmam que, apesar do conhecimento científico continuar a ser gerado por experimentos tradicionais (*in vivo* e *in vitro*) são estudadas novas modalidades de experimentos científicos: *in virtuo* e *in silico* [TRA03] Apud [MAT09], fazendo com que os objetos de análise dos experimentos sejam usualmente processados por simulações computacionais, permitindo observar o mundo real por meio de simulações em ambientes virtuais.

Apesar disso, Mattoso et al. [MAT09] também afirmam que o cenário atual remete aos primórdios da computação, pois as pesquisas ainda dependem da capacidade individual dos cientistas para o encadeamento das etapas e programas necessários para a execução de experimentos, sendo sujeito a falhas e, muitas vezes, improdutivos, especialmente em se tratando de experimentos complexos, envolvendo muitos programas e grandes quantidades de dados. Em função disso, os sistemas de gerenciamento de *workflows* científicos passaram a ser utilizados.

A implementação de processos antes manuais minimizou o problema, mas não o resolveu por completo. Os tempos envolvidos ainda são consideráveis, justificando novos estudos na busca de otimização. Diferentemente do *workflow* científico desenvolvido por Karina Machado, em [MAC07a], que apresenta características puramente sequenciais, o trabalho desenvolvido nesta Tese tem o propósito de possibilitar a execução de experimentos em paralelo e reduzir a quantidade total de execuções de experimentos, utilizando-se dos conceitos de padrão de dados e de adaptação de instâncias em execução por meio da otimização dos recursos computacionais. O Capítulo 4 apresenta padrões de dados e de fluxos utilizados como norteadores para o padrão desenvolvido no Capítulo 5.

4 PADRÕES DE WORKFLOWS

Russel et al. [RUS04] afirmam que existem vários conceitos utilizados na representação e manipulação de dados em sistemas de *workflow*. Esses conceitos não apenas definem como os dados devem ser armazenados, mas sua aplicação em processos de negócio e a interação com outros sistemas e ambientes. Os autores apresentam uma série de padrões de dados, os quais objetivam capturar as diferentes formas de representação e utilização dos dados sobre *workflows*. Para os autores, uma vantagem significativa na abordagem baseada em padrões está em servir como base de comparação entre ferramentas diferentes. Buscando aprimorar essa base, Russel et al. [RUS04] estenderam um estudo realizado por Jablonski e Bussler [JAB96], os quais identificaram 40 padrões de dados e os analisaram sobre seis produtos de *workflow* de mercado da época. Esses padrões são apresentados nas próximas seções. Além dos padrões de dados, este capítulo também apresenta os padrões de controle de fluxo utilizados por Nardi em [NAR09] para a definição do padrão Junção Combinada. A notação diagramática utilizada para representar os padrões de dados foi definida por Russel et al., em [RUS04], enquanto a notação utilizada para a representação dos padrões de controle de fluxo é a da Rede de Petri Colorida, já detalhada no Capítulo 2. Um dos principais motivos para o estudo destes padrões está na identificação da existência de algum padrão que pudesse ser capaz de atender à necessidade de áreas como a Bioinformática: manipular grandes volumes de dados, com características semelhantes, na menor quantidade de tempo possível.

4.1 Padrões de Dados

Russel et al. [RUS04] classificam soluções conforme a perspectiva dos dados e suas características. Dentre essas perspectivas estão:

- *Visibilidade dos dados*: dados podem ser vistos pelos diversos componentes de um processo de *workflow*.
- *Interação dos dados:* dados se comunicam entre atividades de um *workflow*.

- *Transferência de dados*: considera a transferência entre componentes do *workflow* e descreve os diversos mecanismos pelos quais os dados podem ser enviados de uma interface para um componente do *workflow*.
- *Roteamento baseado em dados*: de como os tipos de dados podem influenciar as operações e outros aspectos do *workflow*, principalmente quando se analisa fluxos de controle.

4.1.1 Visibilidade dos dados

Para Russel et al. [RUS 04 e RUS 05], dentro do contexto de *workflow*, existem formas distintas para a definição e utilização dos dados. Tipicamente, dados individuais são utilizados em uma estrutura específica de *workflow*, por exemplo, uma tarefa ou um bloco, e essas estruturas definem o escopo de acesso aos dados, especificando, dessa forma, como os dados podem ser utilizados: para buscar informações de produção, para gerenciar o controle aos dados ou a comunicação com o ambiente externo. Assim, os autores identificam oito padrões que manipulam a visibilidade dos dados.

4.1.1.1 Padrão 1. Dados de Tarefas (*Task Data*)

Os dados podem ser definidos pelas tarefas que são acessíveis somente dentro do contexto da execução individual das instâncias dessas tarefas. O objetivo é tornar possível operações locais em nível de tarefa. Geralmente, esses dados são utilizados para prover armazenamento de informações durante a execução da tarefa ou resultados intermediários para manipulação de dados produzidos. A Figura 14 ilustra a declaração de um dado em uma tarefa (variável X na tarefa B) e o escopo no qual esse dado pode ser utilizado. Há uma distinção na execução da tarefa B para cada instância em execução, podendo a variável X assumir diferentes valores.

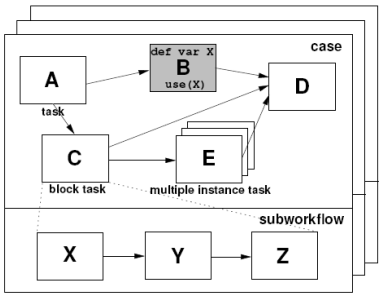

Figura 14 – Visibilidade dos Dados em Nível de Tarefa [RUS04]

4.1.1.2 Padrão 2. Dados de Blocos

Tarefas de blocos são tarefas que podem ser descritas como subprocessos. São capazes de definir dados acessíveis pelos componentes do subprocesso correspondente. Dados definidos na tarefa do processo principal são utilizados no subprocesso, conforme ilustra a Figura 15.

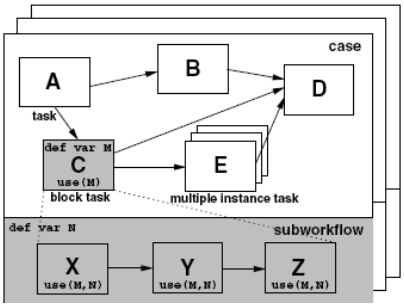

Figura 15 – Visibilidade dos Dados em Nível de Bloco [RUS04, RUS05]

4.1.1.3 Padrão 3. Dados por Escopo

Os dados podem ser definidos para serem acessíveis por um subconjunto de tarefas, conforme a aplicação. Esse padrão é aplicado em casos onde muitas tarefas direcionam suas ações para dados ou conjuntos de dados comuns. A Figura 16 exemplifica esse padrão.

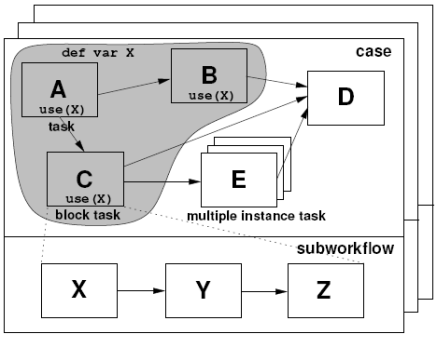

Figura 16 – Visibilidade dos Dados em Nível de Escopo [RUS04]

4.1.1.4 Padrão 4. Dados de Múltiplas Instâncias

Tarefas habilitadas a executarem múltiplas vezes sobre um processo podem possuir dados que sejam definidos sobre uma execução individual de uma instância sobre esse processo. Existem três cenários sobre os quais uma tarefa poderia ser executada mais de uma vez, ilustrados na Figura 17: (a) onde uma tarefa é designada como sendo tarefa de múltiplas instâncias e, uma vez habilitada, múltiplas instâncias poderiam ser iniciadas simultaneamente; (b) onde diferentes tarefas em um *workflow* compartilham a mesma implementação; (c) onde uma tarefa pode receber várias entradas durante a execução de um *workflow*.

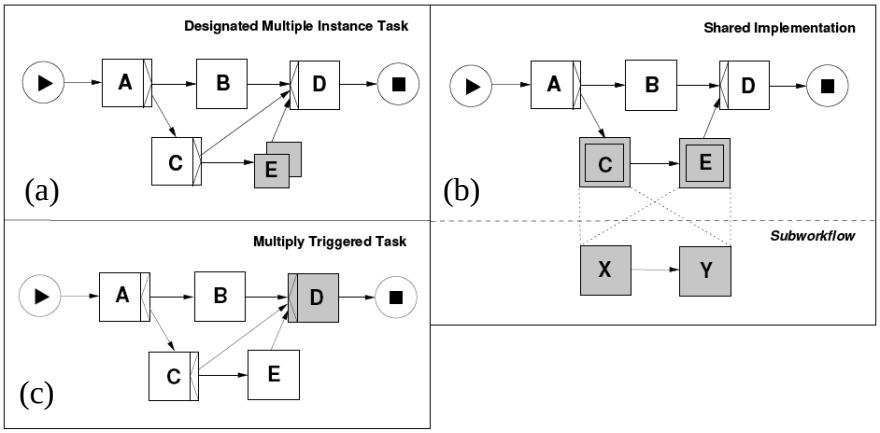

Figura 17 – Visibilidade dos Dados em Múltiplas Instâncias [RUS04]

4.1.1.5 Padrão 5. Dados de Casos

Dados são definidos para uma instância específica ou para um caso de um *workflow*. Dessa forma, podem ser acessados por todos os componentes do *workflow* durante a execução desse caso e os dados são gerenciados como variáveis globais, durante a execução. A Figura 18 representa a utilização de um dado sobre um caso.

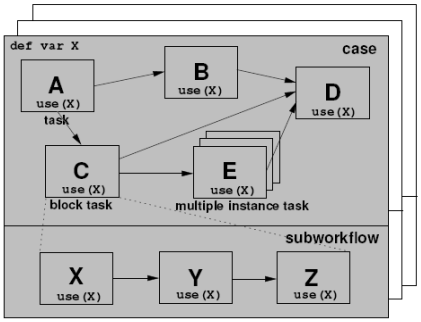

Figura 18 – Visibilidade dos Dados em nível de Caso [RUS04]

4.1.1.6 Padrão 6. Dados de Pastas

Nesse padrão os dados podem der definidos como acessíveis por múltiplos casos em uma base seletiva. Por exemplo: todas as instâncias que passam por uma determinada tarefa podem acessar o valor de um dado específico, possibilitando, dessa forma, o compartilhamento de dados entre instâncias de tarefas de diferentes casos de *workflow*. A Figura 19 ilustra o funcionamento desse padrão.

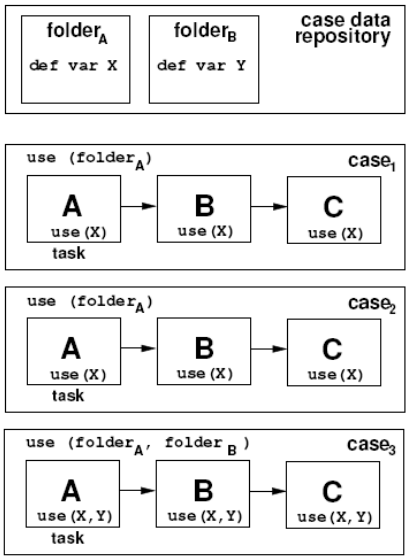

Figura 19– Visibilidade dos Dados em nível de Pastas [RUS04]

4.1.1.7 Padrão 7. Dados de *Workflows*

Nesse padrão os dados são acessados por todos os componentes do *workflow*, em cada um dos seus casos, e são controlados pelo sistema de *workflow*. A Figura 20 ilustra esse padrão. Diferente do padrão 5, nesse padrão os dados são acessíveis por todos os casos do *Workflow*.

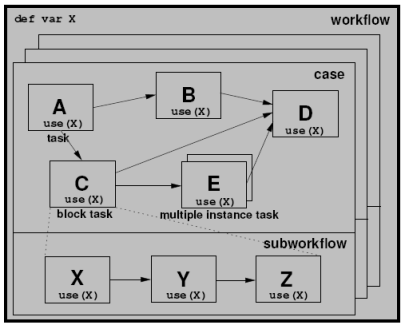

Figura 20 – Visibilidade dos Dados em nível de *Workflow* [RUS04]

4.1.1.8 Padrão 8. Dados de Ambiente

Nesse padrão os dados estão em um ambiente externo, acessados pelos componentes do *workflow* durante a execução. A Figura 21 representa esse padrão.

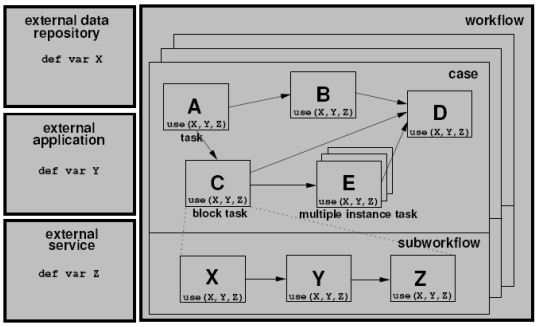

Figura 21 – Visibilidade dos Dados em nível de Ambiente

4.1.2 Interação dos dados

Para Russel et al. [RUS 04 e RUS 05] os padrões de interação capturam as diferentes possibilidades sobre as quais os dados podem ser executados entre os componentes de um *workflow* e como as características desses componentes podem influenciar na forma como esses dados são executados. Um dos principais objetivos é distinguir a interação entre componentes do próprio *workflow* e de alguns recursos utilizados do ambiente externo. Os autores identificaram 18 padrões de interação onde seis desses padrões envolvem apenas componentes internos ao *workflow* e os 12 demais envolvem interação entre os componentes internos e o ambiente externo.

4.1.2.1 Padrão 9. Interação de Dados – Tarefa a Tarefa

É a habilidade de comunicação dos dados entre uma instância de tarefa e outra em um mesmo caso de *workflow*. É a possibilidade de se transmitir dados utilizados por diferentes tarefas quando requeridos. Três possibilidades subsidiam esse padrão, ilustradas na Figura 22: (a) canais de dados e controle integrados: quando dados e fluxos são encaminhados simultaneamente entre tarefas utilizando o mesmo canal; (b) canais de dados distintos: quando o dado é passado entre tarefas do *workflow* por um canal explícito, distinto do controle do processo; (c) armazenamento global de dados: quando as tarefas compartilham os mesmos dados, por compartilhamento global de informações.

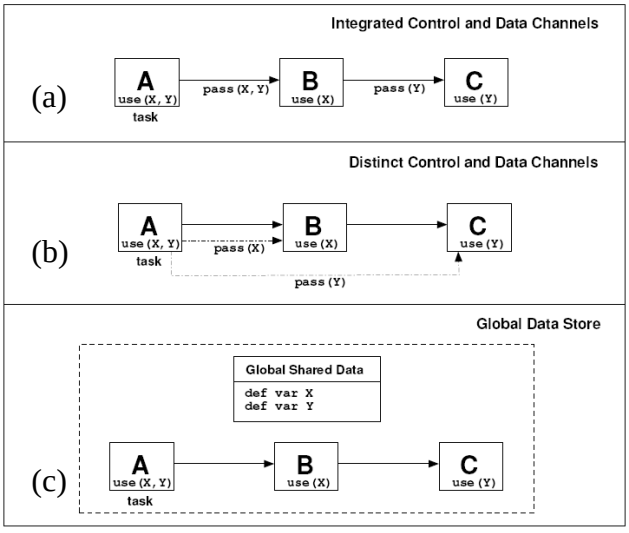

Figura 22 – Interação dos Dados: Abordagens Tarefa a Tarefa [RUS04, RUS05]

4.1.2.2 Padrão 10. Interação de Dados – Bloco de Tarefas para Subworkflow

É a habilidade de passar dados de uma instância de um bloco de tarefas para seu subworkflow correspondente. Esse padrão possui abordagens possíveis para implementação, ilustradas na Figura 23: (a) passagem de dados implícitos: os dados passados pelo bloco são

imediatamente acessíveis por todas as tarefas do subworkflow correspondente e não há a necessidade de passagem explícita de parâmetros; (b) passagem de dados explícita via parâmetros: os dados devem ser especificados como parâmetros, sendo encaminhados para o subworkflow; (c) passagem de dados explícita via canal de dados: os dados são passados especificamente via canal de dados para todas as tarefas do subworkflow que requisitam a informação.

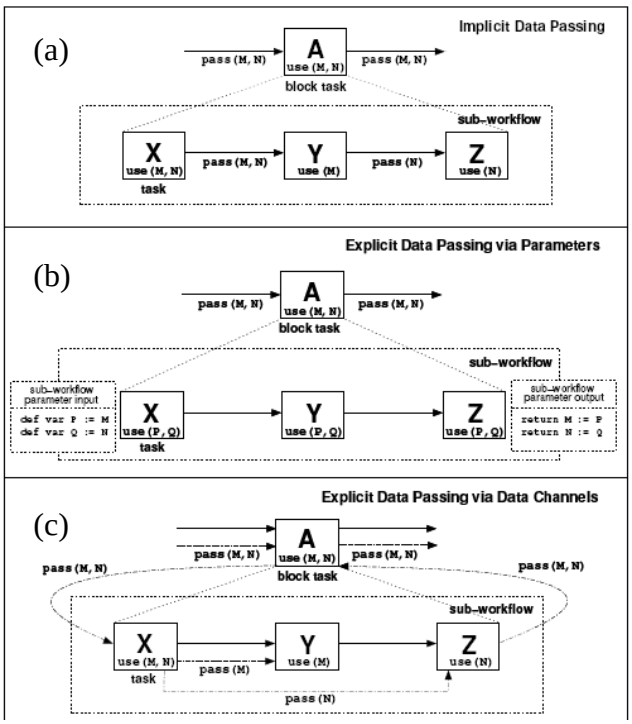

Figura 23 – Interação dos Dados: Abordagens Blocos de Tarefas para Subworkflow [RUS04]

4.1.2.3 Padrão 11. Interação de Dados – Subworkflow para Bloco de Tarefas

É a habilidade de passar dados de um subworkflow para seu bloco de tarefas correspondente. Esse padrão identifica o processo inverso abordado no padrão 10.

4.1.2.4 Padrão 12. Interação de Dados – Tarefas de Múltiplas Instâncias

É a habilidade de passar dados de uma instância de tarefa para tarefas que suportem a execução de múltiplas instâncias. Pode envolver a passagem de dados para todas as instâncias da tarefa seguinte ou conforme critérios predefinidos. Existem três abordagens, apresentadas na Figura 24: (a) dados da instância específica são passados por valor para as tarefas seguintes; (b) dados da instância específica são passados como referência para as tarefas seguintes; (c) dados compartilhados são passados por referência: permite que instâncias de tarefas acessem os mesmos dados, sem realizar controle de concorrência.

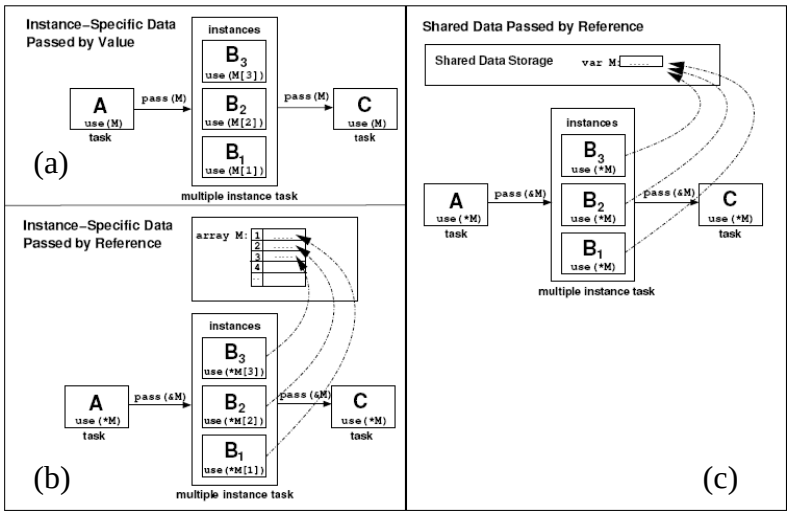

Figura 24 – Interação dos Dados: Tarefas de Múltiplas Instâncias [RUS04, RUS05]

4.1.2.5 Padrão 13. Interação de Dados – de Tarefas de Múltiplas Instâncias

É a habilidade de passar dados de uma tarefa que suporte múltiplas instâncias para uma tarefa seguinte. Cada execução de uma tarefa que suporte múltiplas instâncias efetivamente é executada de forma independente de outras instâncias e deve passar dados visando à conclusão das tarefas seguintes. Assim, os dados gerados, como saída da execução de uma instância de tarefa, devem ser agregados às tarefas seguintes.

4.1.2.6 Padrão 14. Interação de Dados – Casos para Casos

É a passagem de dados durante a execução de um caso para outro caso também em execução. Alternativamente é possível acessar a mesma saída indiretamente por meio do armazenamento dessas informações em uma estrutura de armazenamento compartilhada. Essa alternativa é representada na Figura 25.

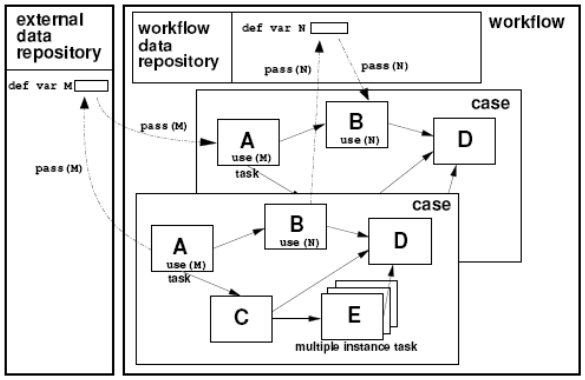

Figura 25 – Interação dos Dados: Casos para Casos [RUS04]

4.1.2.7 Padrão 15. Interação de Dados – Tarefas para Ambiente Externo – Push-Oriented

A habilidade de uma tarefa iniciar a passagem de dados para um recurso ou serviço no ambiente operacional. A Figura 26 ilustra os vários cenários de passagem de dados entre tarefas de *workflow* e o ambiente externo. Existem duas categorias principais que subsidiam a

implementação desse tipo de interação: (i) mecanismo de integração explícito: onde o sistema de *workflow* provê construtores específicos para passagem de dados ao ambiente externo; (ii) mecanismo de integração implícito: onde a passagem dos dados ocorre implicitamente dentro de implementações que fazem chamadas nos processos do *workflow* e não são suportadas diretamente pelo ambiente.

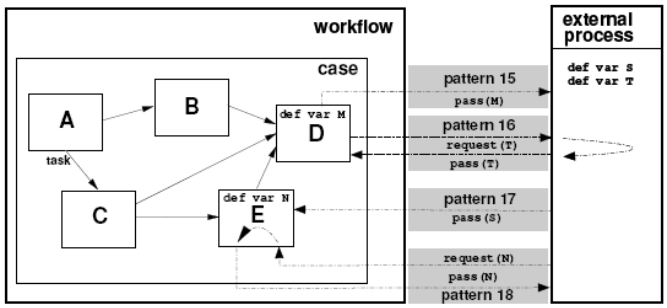

Figura 26 – Interação dos Dados: Tarefas para Ambiente Externo[RUS04]

4.1.2.8 Padrão 16. Interação de Dados – Ambiente Externo para Tarefas – *Pull-Oriented*

A habilidade de uma tarefa de *workflow* requisitar dados de recursos ou serviços de um ambiente operacional. Pode envolver o acesso aos dados de um repositório, por exemplo. A Figura 26 também ilustra esse padrão.

4.1.2.9 Padrão 17. Interação de Dados – Ambiente Externo para Tarefas – *Push-Oriented*

A habilidade de uma tarefa de *workflow* receber e utilizar dados passados de serviços e recursos do sistema operacional, sem um planejamento prévio. A possibilidade das tarefas receberem novos itens de dados assim que estiverem disponíveis, sem a necessidade de requisição. A Figura 26 também ilustra esse padrão.

4.1.2.10 Padrão 18. Interação de Dados – Tarefas para Ambiente Externo – Pull-Oriented

A habilidade de uma tarefa de *workflow* receber e responder a requisições de dados para serviços e recursos no sistema operacional. Essa habilidade pode ser manipulada de três formas: (i) o *workflow* provê formas de acessar dados de instâncias de tarefas do ambiente externo; (ii) durante a execução, as instâncias de tarefas publicam valores dos dados em locais conhecidos, por exemplo, uma base de dados; (iii) as instâncias de tarefas incorporam facilidades aos serviços de requisição de dados aos processos externos. A Figura 26 também ilustra esse padrão.

4.1.2.11 Padrão 19. Interação de Dados – Caso para Ambiente Externo – Push-Oriented

A habilidade de um caso de *workflow* iniciar a passagem de dados para um recurso ou serviço no sistema operacional. Esse padrão é análogo ao padrão 15, exceto pela utilização de casos e não de uma única tarefa. A Figura 27 ilustra esse padrão.

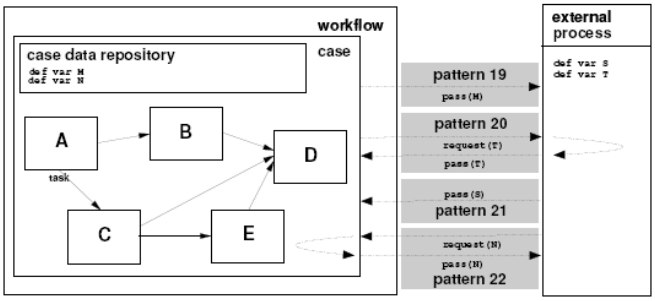

Figura 27 – Interação dos Dados entre Casos e o Ambiente Operacional [RUS04]

4.1.2.12 Padrão 20. Interação de Dados – Ambiente Externo para Caso– Pull-Oriented

A habilidade de um caso de *workflow* requisitar dados de recursos ou serviços do sistema operacional. A Figura 27 também ilustra esse padrão.

4.1.2.13 Padrão 21. Interação de Dados – Ambiente Externo para Caso– Push-Oriented

A habilidade de um caso de *workflow* aceitar dados passados de recursos ou serviços do sistema operacional. Existem duas formas de implementação desse padrão: (i) valores dos dados podem ser especificados na inicialização de um caso específico; (ii) o *workflow* pode providenciar formas de habilitar atualização de dados durante a execução de um caso. A Figura 27 também ilustra esse padrão.

4.1.2.14 Padrão 22. Interação de Dados – Caso para Ambiente Externo – Pull-Oriented

A habilidade de um caso de *workflow* responder a requisições de dados passados para recursos ou serviços do sistema operacional. A Figura 27 também ilustra esse padrão.

4.1.2.15 Padrão 23. Interação de Dados – *Workflow* para Ambiente Externo – Push-Oriented

A habilidade de um *workflow* enviar dados para recursos ou serviços do sistema operacional. Interessante quando as aplicações externas necessitam de informações como: casos de sucesso, recursos utilizados etc. A Figura 28 ilustra esse padrão.

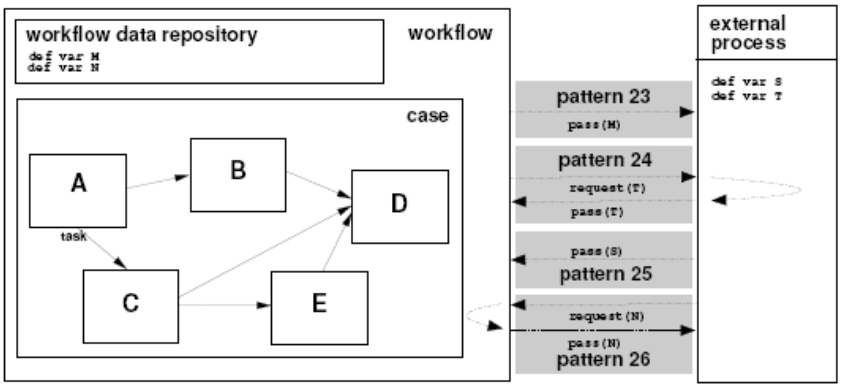

Figura 28 – Interação dos Dados entre um Sistema de *Workflow* e o Ambiente Operacional [RUS04]

4.1.2.16 Padrão 24. Interação de Dados – Ambiente Externo para *Workflow* – Pull-Oriented

A habilidade de um *workflow* requisitar dados no nível de *workflow* de aplicações externas. A Figura 28 também ilustra esse padrão.

4.1.2.17 Padrão 25. Interação de Dados – Ambiente Externo para *Workflow* – Push-Oriented

A habilidade de serviços ou recursos, no ambiente operacional externo, enviar dados para um processo de *workflow*. O objetivo desse padrão é suportar aplicações independentes de ferramenta de *workflow*, com a possibilidade de criar ou atualizar dados nessas ferramentas. Para isso pode-se proceder das seguintes formas: (i) os dados são passados para a ferramenta de *workflow* por linha de comando, por exemplo, no momento que a ferramenta é inicializada (assume-se assim, que a aplicação externa inicializou a ferramenta de *workflow*); (ii) a aplicação externa inicializa a importação de dados pela ferramenta de *workflow*; (iii) a aplicação externa utiliza APIs para acessar conjuntos de dados na ferramenta de *workflow*. A Figura 28 também ilustra esse padrão.

4.1.2.18 Padrão 26. Interação de Dados – *Workflow* para Ambiente Externo – Pull-Oriented

A habilidade de uma ferramenta de *workflow* manipular requisições de dados para aplicações externas. É importante destacar que nesse padrão a requisição é iniciada pela aplicação externa. A implementação desse padrão pode ser realizada das seguintes formas: (i) aplicações externas podem utilizar recursos disponibilizados pelas ferramentas de *workflow* que exportem dados para arquivos, podendo ser acessados pelas aplicações; (ii) aplicações externas podem utilizar APIs disponibilizadas pelas próprias ferramentas de *workflow* que acessem os dados requisitados. A Figura 28 também ilustra esse padrão.

4.1.3 Transferência dos dados

Para Russel et al. [RUS 04 e RUS 05] os padrões de transferência de dados baseiam-se em como a troca de informações ocorre entre componentes do *workflow*. Esses padrões apresentam-se como uma extensão dos já apresentados na seção anterior e objetivam capturar as diversas formas pelas quais os dados são manipulados por meio das interfaces de componentes do *workflow*, envolvendo, por exemplo, fatores como a possibilidade de acesso exclusivo ou compartilhado a uma determinada informação. Esses são os padrões 27 a 33.

4.1.3.1 Padrão 27. Transferência de Dados por Valor – Entrada

A habilidade de um componente do *workflow* receber dados de entrada por valor de outro componente. A Figura 29 ilustra esse padrão.

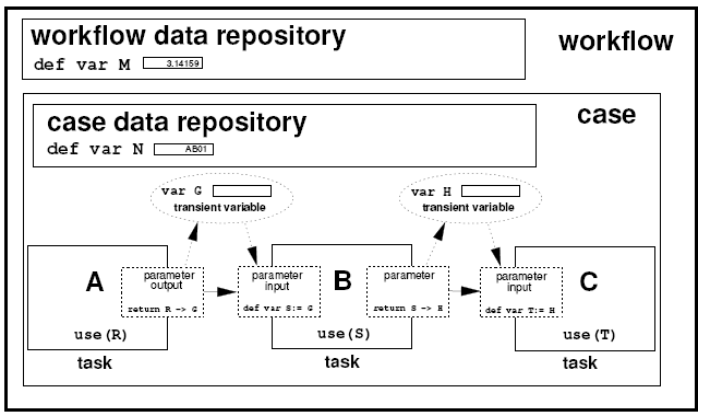

Figura 29 – Transferência de Dados por Valor [RUS04]

4.1.3.2 Padrão 28. Transferência de Dados por Valor – Saída

A habilidade de um componente do *workflow* passar dados por valor para outro componente. A Figura 29 também ilustra esse padrão.

4.1.3.3 Padrão 29. Transferência de Dados – Copy In/Copy Out

A habilidade de um componente do *workflow* copiar os valores de um conjunto de dados para dentro de um repositório de dados no início da execução e copiar o valor final desses dados após a execução ser completada. A Figura 30 ilustra esse padrão.

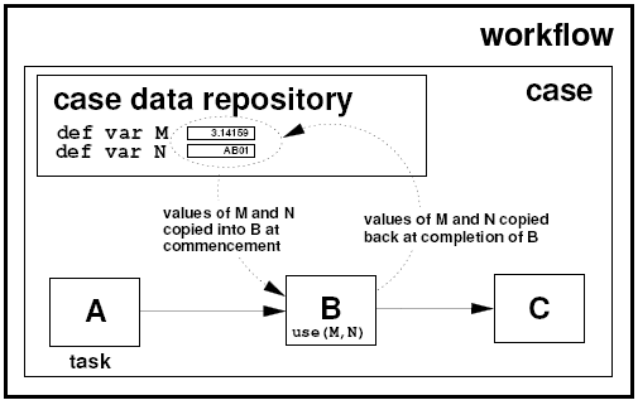

Figura 30 – Transferência de Dados – Copy In/Copy Out [RUS04]

4.1.3.4 Padrão 30. Transferência de Dados por Referência – Desbloqueado

A habilidade de comunicação de dados entre componentes de *workflow* pela utilização de um local de acesso compartilhado aos dados. Não existem restrições de concorrência aplicadas aos dados. A Figura 31 ilustra esse padrão.

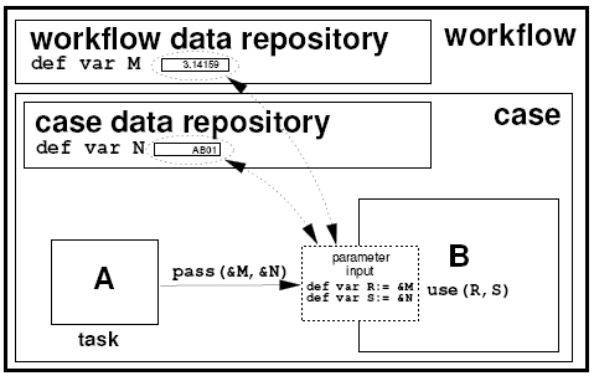

Figura 31 – Transferência de Dados por Referência – Desbloqueado [RUS04]

4.1.3.5 Padrão 31. Transferência de Dados por Referência – Com Bloqueio

A habilidade de comunicação de dados entre componentes de *workflow* pela utilização de um local de acesso compartilhado aos dados. Restrições de concorrência são aplicadas por meio de privilégios de somente leitura ou acesso dedicado. Essa abordagem estende o padrão 30, bloqueando os dados para leitura ou escrita.

4.1.3.6 Padrão 32. Transformação de Dados – entrada

A habilidade de aplicar uma função de transformação sobre um dado antes de ser encaminhado a um componente do *workflow*. A Figura 32 ilustra esse padrão.

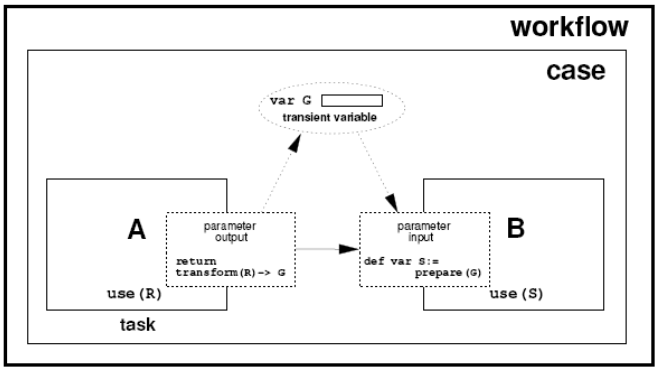

Figura 32 – Transformação de Dados – Entrada e Saída [RUS04]

4.1.3.7 Padrão 33. Transformação de Dados – saída

A habilidade de aplicar uma função de transformação sobre um dado imediatamente antes de sair de um componente do *workflow*. A Figura 32 também ilustra esse padrão.

4.1.4 Roteamento baseado em dados

Para Russel et al. [RUS 04] os padrões de roteamento baseado em dados estudam as diversas formas por meio das quais os dados podem interagir com outras perspectivas e influenciar a operação de um *workflow* como um todo. Esses são os padrões 34 a 40.

4.1.4.1 Padrão 34. Pré-condição para Tarefa – Existência de Dados

Pré-condições com base em dados podem ser especificadas por tarefas que definem sua execução pela presença de dados em tempo de execução. A Figura 33 ilustra esse padrão.

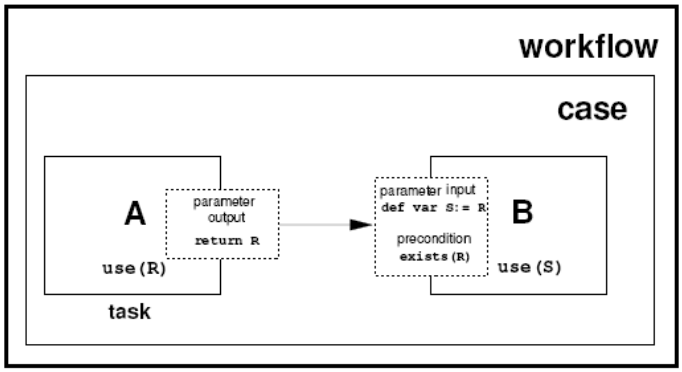

Figura 33 – Pré-condição para Tarefa com base na Existência de Dados [RUS04]

4.1.4.2 Padrão 35. Pré-condição para Tarefa – Valor de Dados

Pré-condições com base em dados podem ser especificadas por tarefas que definem sua execução pela presença de um determinado valor para o dado, especificado em tempo de execução.

4.1.4.3 Padrão 36. Pós-Condição para Tarefa – Existência de Dados

Pós-condições com base em dados podem ser especificadas por tarefas que definem sua execução pela existência de parâmetros específicos em tempo de execução. A Figura 34 ilustra esse padrão, onde a pós-condição de uma tarefa efetivamente estabelece um loop de controle implícito até que ela seja satisfeita.

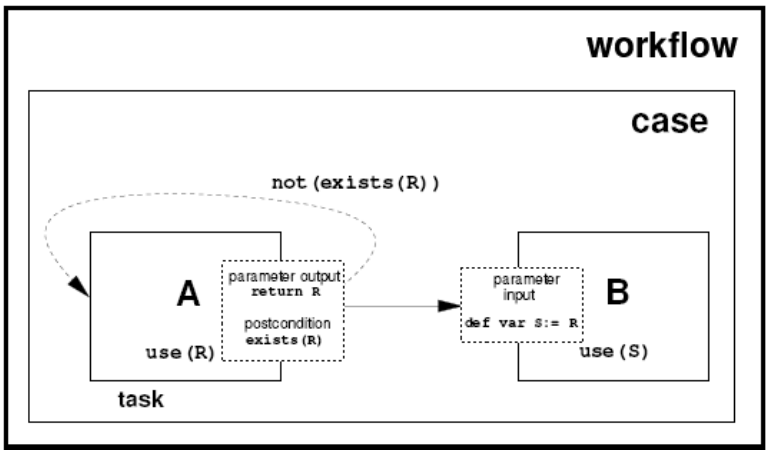

Figura 34 – Pós-condição para Tarefa com base na Existência de Dados [RUS04]

4.1.4.4 Padrão 37. Pós-Condição para Tarefa – Valor de Dados

Pós-condições com base em dados podem ser especificadas por tarefas que definem sua execução por valores de parâmetros específicos em tempo de execução.

4.1.4.5 Padrão 38. Disparo de Tarefas com Base em Eventos

A habilidade de um evento externo inicializar uma tarefa. A Figura 35 ilustra esse padrão.

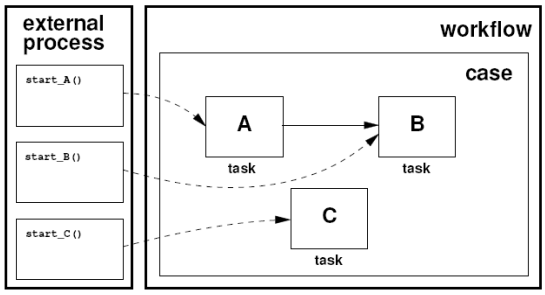

Figura 35 – Disparo de Tarefas com base em Eventos Externos [RUS04]

4.1.4.6 Padrão 39. Disparo de Tarefas com Base em Dados

A habilidade de disparar uma tarefa específica quando a avaliação de uma expressão de dados do *workflow* é verdadeira. A Figura 36 ilustra esse padrão.

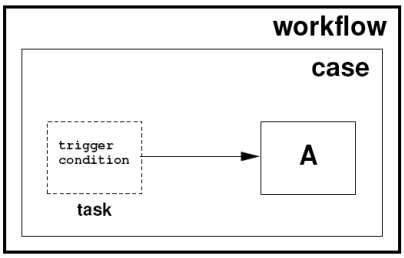

Figura 36 – Disparo de Tarefas com base em Dados [RUS04]

4.1.4.7 Padrão 40. Roteamento Baseado em Dados

A habilidade de alterar o fluxo dentro de um caso de *workflow* após a análise de expressões baseadas em dados. A Figura 37 ilustra esse padrão. Esse padrão serve como uma agregação dos dois maiores padrões de controle de fluxo que dependem de dados: (i) escolha exclusiva, onde o fluxo de controle é passado para uma das muitas tarefas seguintes dependendo da saída de uma decisão ou do valor de uma expressão; (ii) escolha múltipla, onde dependendo da saída de uma decisão ou do valor de uma expressão, o fluxo de controle é passado para várias tarefas seguintes.

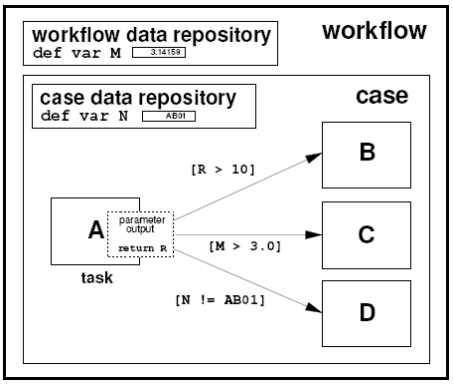

Figura 37 – Roteamento Baseado em Dados [RUS04]

4.2 Padrões para Controle de Fluxo

Os padrões para Controle de Fluxo (*Workflow* Control-Flow Patterns – WCP) compõem um conjunto proposto em Aalst et al. [AAL03a] de vinte padrões encontrados na execução das etapas de um *workflow*. Russell et al. [RUS06] identificaram diversas especializações desses padrões, completando um total de quarenta e três. Algumas das categorias:

- Padrões de Controle de Fluxo Básicos: essa classe de padrões captura os aspectos elementares de controle de fluxo.
- Padrões Avançados para Ramificações e Sincronizações: para Nardi [NAR09] esses padrões detalham diversas possibilidades para execução de ramificações como discriminadores e junções parciais, incluindo tratamento de atividades pendentes.
- Padrões para Múltiplas Instâncias: descrevem situações onde são executadas, sobre uma mesma atividade, diversas instâncias. Conforme Nardi [NAR09], esse é um caso específico de padrões para ramificação e sincronização. Aqui, todas as atividades são instâncias de uma mesma atividade.
- Padrões Baseados em Estado: de acordo com Nardi [NAR09] “quando a relação entre atividades a serem executadas dependerem de estado, o controle do fluxo deve prever situações de concorrência e ordem de execução”.
- Padrões para Cancelamento e Conclusão Forçada: muitos dos padrões citados anteriormente utilizam o conceito de cancelamento de atividade ou tratamento de exceções por meio do conceito de cancelamento. Esse conjunto de padrões trata o cancelamento isoladamente. Conforme Nardi [NAR09] o Padrão de Conclusão Forçada permite que, a partir de informações externas, as atividades pendentes sejam concluídas.

No contexto de padrões de controle de fluxo, esta Tese tem como uma de suas bases o trabalho desenvolvido por Nardi [NAR09]. Em função disso, são apresentados os padrões de controle de fluxo estudados pelo autor. Nardi descreveu um padrão de divisão paralela e dez padrões de sincronização apresentados por Russel et al. [RUS06] e definiu o Padrão Junção Combinada, como resultado da integração dos padrões estudados. Conforme Nardi [NAR09], dos quarenta e três padrões descritos por Russell et al. [RUS06], vinte e sete se relacionam a ramificação, sincronização ou múltiplas instâncias, podendo se aproveitar da execução em grades,

utilizado no trabalho do autor. Nardi se propôs a tratar o padrão “(WCP-2) Divisão Paralela”, e os seguintes padrões de sincronização:

- WCP-3 Sincronização;
- WCP-9 Discriminador Estruturado;
- WCP-28 Discriminador com Bloqueio;
- WCP-29 Discriminador com Cancelamento;
- WCP-30 Junção Parcial Estruturada;
- WCP-31 Junção Parcial com Bloqueio;
- WCP-32 Junção Parcial com Cancelamento;
- WCP-34 Junção Parcial Estática para Múltiplas Instâncias;
- WCP-35 Cancelamento de Junção Parcial de Múltiplas Instâncias;
- WCP-36 Junção Parcial Dinâmica para Múltiplas Instâncias.

Os Discriminadores, conforme Nardi [NAR09] caracterizam-se pela produção da saída quando uma das atividades em paralelo tiver sido concluída. O autor também apresenta o conceito de junções que, semelhantes aos discriminadores, são uma forma mais genérica. Nas junções a saída é produzida quando uma determinada quantidade n tiver sido concluída, sendo que n é menor ou igual que a quantidade total de entradas possíveis, enquanto os discriminadores são junções parciais onde n=1. O autor também apresenta que padrões de Múltiplas Instâncias são aqueles onde as atividades a serem executadas em paralelo são instâncias de uma mesma atividade, quando as instâncias puderem ser executadas independentemente umas das outras. A notação diagramática utilizada para a representação dos padrões de controle de fluxo é a da Rede de Petri Colorida.

4.2.1 WCP-2 Divisão Paralela

Aalst [AAL07] e Russel [RUS06] definem esse padrão como sendo a existência de uma divisão em dois ou mais ramos paralelos, cada um deles executando concorrentemente. Esses ramos podem ou não ser sincronizados, novamente, em algum ponto do futuro. Para Nardi [NAR09] esse é o padrão que explicita a possibilidade de execução em grade, pois as saídas desse padrão são as entradas para os padrões de sincronização, descritos nas próximas seções. A Figura 38 representa esse padrão.

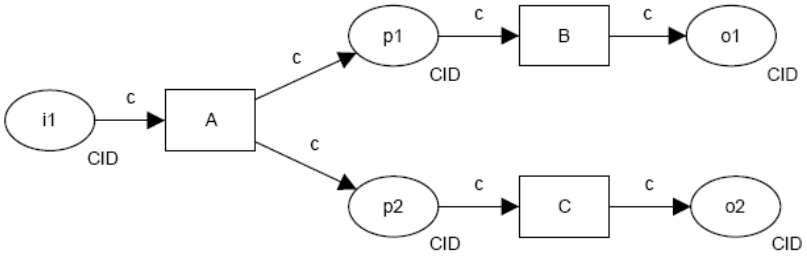

Figura 38 – Padrão Divisão Paralela [RUS06]

4.2.2 WCP-3 Sincronização

Aalst [AAL07] e Russel [RUS06] definem esse padrão como sendo a convergência de dois ou mais ramos (caminhos de execução paralela) em um ramo único por onde passa o controle para as próximas ramificações e caminhos quando todos os demais ramos tiverem sido habilitados (E lógico). Provê um mecanismo de convergência de duas ou mais ramificações paralelas. Em geral, esses ramos foram criados a partir de divisões paralelas. A Figura 39 representa o padrão Sincronização.

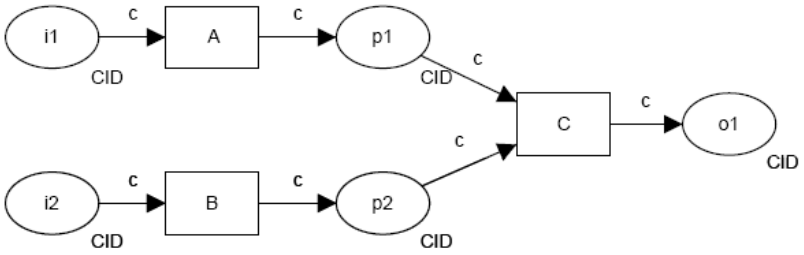

Figura 39 – Padrão Sincronização [RUS06]

4.2.3 WCP-9 Discriminador Estruturado

Conforme Nardi [NAR09], assim como os demais padrões de sincronização, o Discriminador Estruturado trabalha juntamente com o padrão Divisão Paralela WCP-2. Russel et al. [RUS06] afirmam que a convergência de duas ou mais entradas em uma única saída é um exemplo desse padrão. A Figura 40 apresenta esse discriminador.

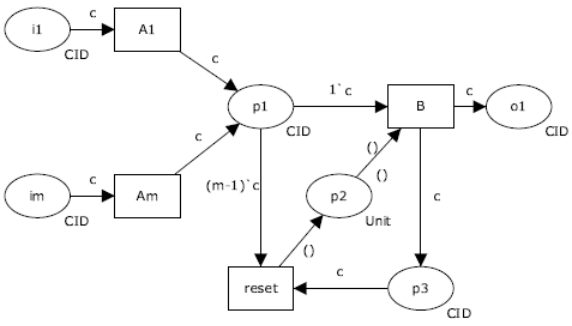

Figura 40 – Padrão Discriminador Estruturado [RUS06]

Nardi [NAR09] afirma que a aplicação que utiliza o padrão “Discriminador Estruturado” é responsável por garantir que uma nova execução no mesmo processo não seja possível até que todas as tarefas tenham sido concluídas. Caso essa característica não seja atendida, uma nova execução da tarefa que tenha sido concluída será possível, com consequências que o padrão não prevê.

4.2.4 WCP-28 Discriminador com Bloqueio

O padrão Discriminador com Bloqueio [RUS06, NAR09] contém elementos para não permitir que uma execução seja possível antes do término da execução anterior, independente da aplicação que o utiliza (diferente do padrão Discriminador Estruturado que depende da aplicação). Trata-se de uma especialização do Padrão Discriminador Estruturado. Aqui, o padrão independe da aplicação no controle de execuções sucessivas. A Figura 41 representa esse padrão.

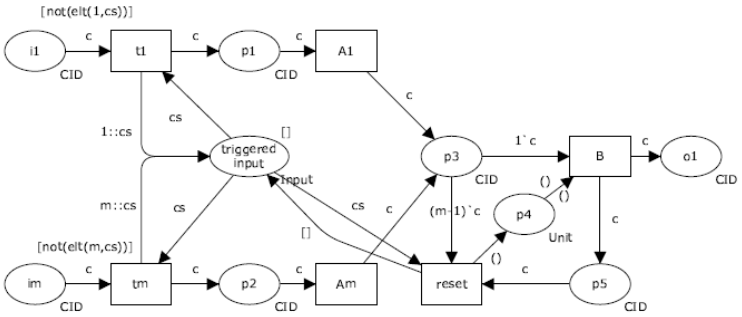

Figura 41 – Padrão Discriminador com Bloqueio [RUS06]

4.2.5 WCP-29 Discriminador com Cancelamento

Para Russel et al. [RUS06] esse padrão provê o cancelamento de entradas que não sejam necessárias. Por exemplo, em situações onde uma entrada seja requisitada, as demais são canceladas para que o processamento seja continuado. Nardi [NAR09] afirma que ao produzir uma saída o Padrão Discriminador Estruturado aguarda que todas as demais atividades sejam concluídas para então poder ser novamente utilizado, mas quando essa espera for desnecessária, o padrão pode ser colocado em prontidão novamente ao sinalizar às demais atividades que seu resultado será descartado, ação realizada pelo Discriminador com Cancelamento. A Figura 42 apresenta esse padrão.

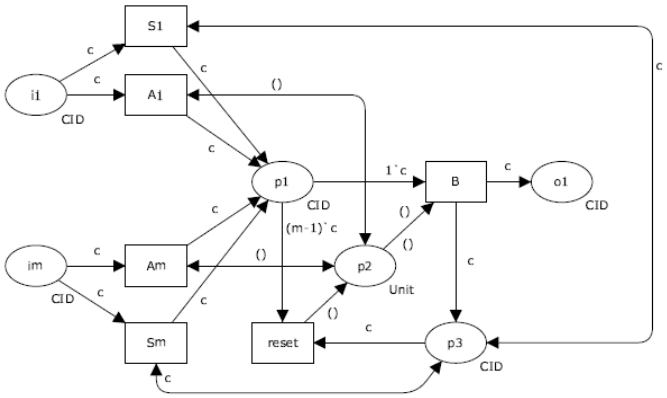

Figura 42 – Padrão Discriminador com Cancelamento [RUS06]

4.2.6 WCP-30 Junção Parcial Estruturada

Nardi [NAR09] afirma que esse padrão é semelhante aos Discriminadores Estruturados, a única diferença é a transição responsável pela junção ser habilitada quando n atividades tiverem sido concluídas. A transição que indica que a atividade foi finalizada permanece sendo habilitada quando as demais atividades forem concluídas. Enquanto nos discriminadores eram (m-1) atividades, nas junções parciais esse número é (m-n). O padrão é representado na Figura 43.

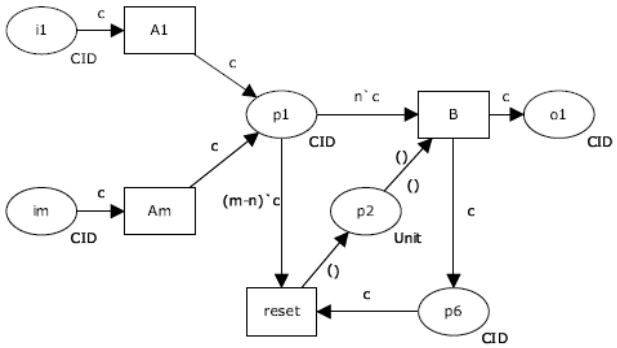

Figura 43 – Padrão Junção Parcial Estruturada [RUS06]

4.2.7 WCP-31 Junção Parcial com Bloqueio

Russel et al. [RUS06] afirmam que o padrão Junção Parcial com Bloqueio é uma variação do padrão Junção Estruturada, que possibilita a execução em ambiente onde são instanciados processos concorrentes. Para Nardi [NAR09], esse padrão trabalha do mesmo modo que os Discriminadores com Bloqueio, tornando o padrão independente da aplicação, no tocante ao tratamento de bloqueios de execução de atividades antes do término da execução anterior. A Figura 44 apresenta esse padrão.

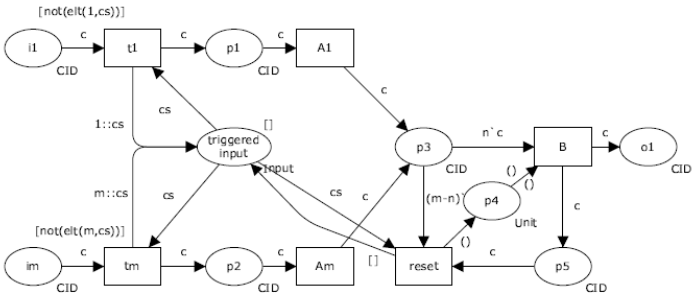

Figura 44 – Padrão Junção Parcial com Bloqueio [RUS06]

4.2.8 WCP-32 Junção Parcial com Cancelamento

Russel et al. [RUS06] afirmam que este padrão prevê que um conjunto de entradas necessitam de sincronização em uma junção, mas somente um subconjunto delas necessita ser finalizado. Para Nardi [NAR09] esse padrão é semelhante aos Discriminadores com Cancelamento, permite a reutilização do padrão após n tarefas terem sido concluídas, cancelando as demais, podendo ser analisado na Figura 45.

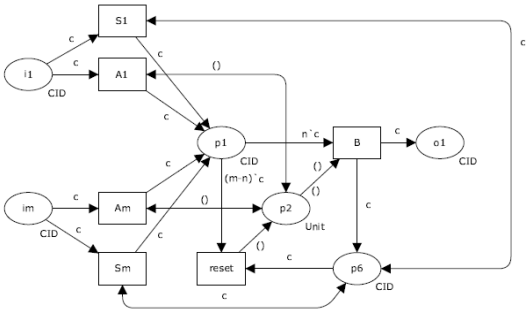

Figura 45 – Padrão Junção Parcial com Cancelamento [RUS06]

4.2.9 WCP-34 Junção Parcial Estática para Múltiplas Instâncias

Em Russel et al. [RUS06] encontra-se que com uma determinada instância do processo, múltiplas instâncias concorrentes de uma atividade podem ser criadas e uma vez que N instâncias dessa atividade forem finalizadas, a próxima atividade é inicializada. Nardi [NAR09] afirma que, embora Russell et al. [RUS06] apresentem este como um novo padrão, trata-se de uma especialização do Padrão Junção Parcial Estruturada, em que as atividades a serem executadas possuem mesmo código, ou seja, são instâncias de uma mesma atividade, possivelmente com parâmetros distintos. A Figura 46 ilustra esse padrão.

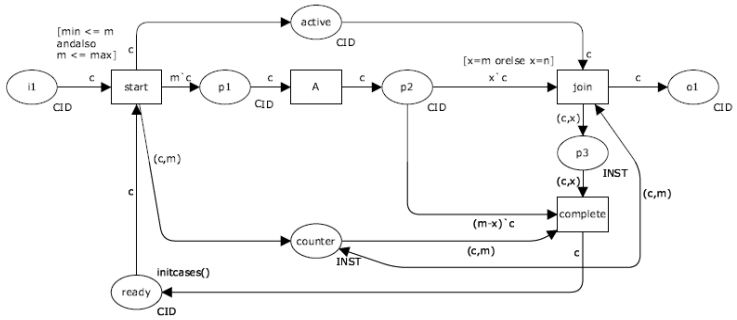

Figura 46 – Padrão Junção Parcial Estática para Múltiplas Instâncias [RUS06]

4.2.10 WCP-35 Junção Parcial de Múltiplas Instâncias com Cancelamento

Em Russel et al. [RUS06] encontra-se que com uma determinada instância do processo, múltiplas instâncias concorrentes de uma atividade podem ser criadas e uma vez que N instâncias dessa atividade forem finalizadas, a próxima atividade é inicializada e as instâncias restantes são canceladas. Nardi [NAR09] afirma que, embora Russell et al. [RUS06] apresentem este como um novo padrão, trata-se de uma especialização do Padrão Junção Parcial com Cancelamento, em que as atividades a serem executadas possuem mesmo código, ou seja, são instâncias de uma mesma atividade, possivelmente com parâmetros distintos. A Figura 47 ilustra este padrão.

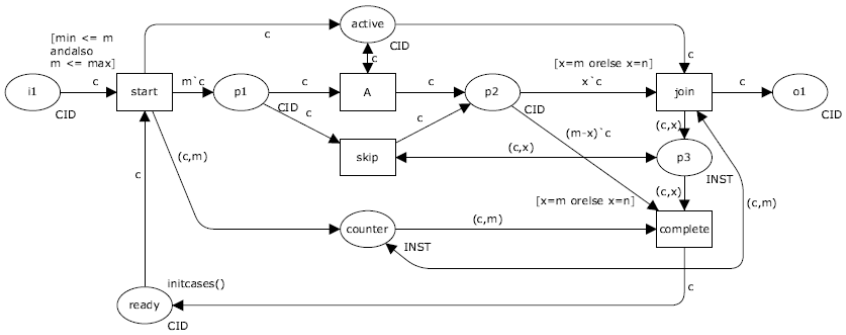

Figura 47 – Padrão Junção Parcial de Múltiplas Instâncias com Cancelamento [RUS06]

4.2.11 WCP-36 Junção Parcial Dinâmica de Múltiplas Instâncias

Com um processo instanciado, múltiplas instâncias concorrentes de uma atividade podem ser criadas. Assim, o número de instâncias pode depender de diversos fatores dentre eles: dados de estado, disponibilidade de recursos, comunicação entre os processos e até mesmo não ser conhecida até que a última instância seja finalizada, uma vez que a execução de uma instância pode gerar a criação de uma nova [RUS06]. Nardi [NAR09] afirma que esse padrão é uma extensão do padrão Junção Parcial Estática de Múltiplas Instâncias, sendo possível executar novas instâncias da atividade em questão, antes que o padrão tenha produzido uma saída. A Figura 48 ilustra esse padrão.

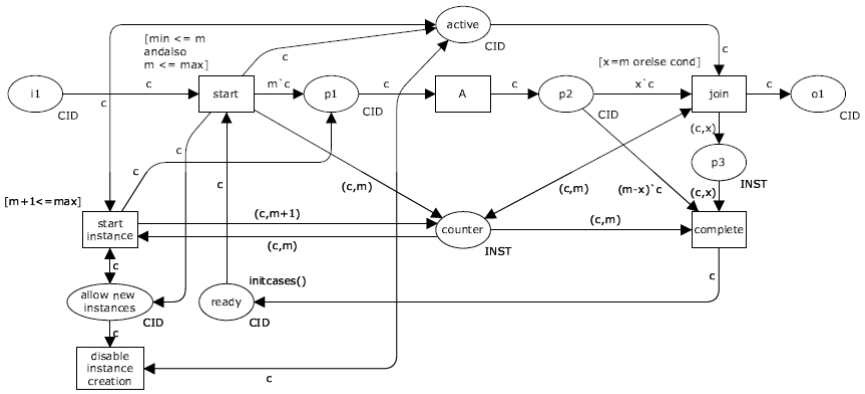

Figura 48 – Padrão Junção Parcial Dinâmica de Múltiplas Instâncias [RUS06]

4.2.12 Padrão Junção Combinada

Nardi [NAR09] apresenta um novo padrão, denominado PJC – Padrão Junção Combinada. A Figura 49 ilustra esse padrão. Esse padrão é uma proposta para a representação e implementação conjunta dos padrões apresentados anteriormente.

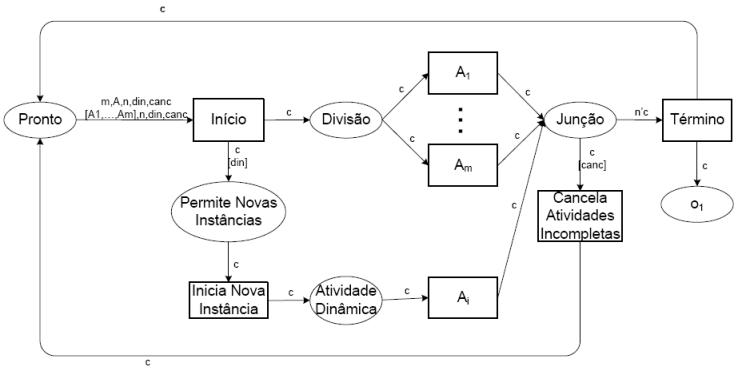

Figura 49 – Padrão Junção Combinada [NAR09]

Nardi apresenta como resultado do seu trabalho uma arquitetura de baixo acoplamento para execução de padrões de controle de fluxo, potencialmente com a utilização de grades. O padrão desenvolvido por Nardi permite a execução de padrões de paralelização em *clusters*, grades ou equipamentos multi-processados, sendo uma base útil para a aplicação do padrão desenvolvido nesta Tese. Além disso, o trabalho desenvolvido pelo autor também oferece infraestrutura extensível para ser complementada com a implementação de outros padrões. O padrão Junção Combinada, também desenvolvido para o contexto de *workflows* científicos, apresenta características de paralelismo, mas não tem como reduzir a quantidade de experimentos a serem processados, por não possuir um tipo de retorno nas execuções que permita a reconfiguração do experimento como um todo, fazendo com que a manipulação de grandes volumes de dados, da área de Bioinformática, ainda seja exaustiva.

4.3 Considerações do Capítulo

A definição de padrões, que representem a manipulação e o fluxo de dados, possibilita a criação de sistemas que gerenciem *workflows* com base em estruturas semelhantes, facilitando a modelagem e aumentando a possibilidade de migração de uma plataforma para outra. Além disso, conforme Nardi [NAR09], essa definição faz com que cientistas de áreas diferentes da computação tenham mais tempo para suas pesquisas, ao invés de ocupá-los na construção da infraestrutura necessária para testar suas teorias ou conceitos. Este capítulo apresentou os padrões de dados e alguns padrões de controle de fluxo. Apesar disso, nenhum dos padrões estudados atende completamente à necessidade de áreas como a Bioinformática, na qual um grande volume de informações é manipulado e encaminhado para processamento. São informações, muitas vezes, semelhantes, podendo-se utilizar dessa semelhança para otimizar o processamento e que resultados preliminares podem auxiliar na definição e recalibragem dos próximos passos. Definem-se, portanto, como características para um padrão aplicado a áreas como a Bioinformática:

- manipulação de múltiplas instâncias;
- visualização dos dados fora do ambiente de *workflow*;
- troca de informações entre diferentes casos de *workflow*;
- passagem de dados para um recurso ou serviço no ambiente operacional;
- passagem de dados do ambiente operacional para o sistema de *workflow*;
- transferência de dados por valor de um componente do *workflow* para outro;
- interpretação de uma pré-condição para a execução de uma tarefa;
- definição de uma pós-condição para a continuidade do processo;
- definição de fluxo com base em condições;
- manipulação de agrupamentos de dados;
- determinação da instanciação de um processo com base em resultados já obtidos;
- determinação da continuidade de um processo com base em resultados já obtidos;
- determinação da alteração de prioridade de uma execução com base em resultados já obtidos;

Nesse contexto, a partir da análise dos padrões de dados e padrões de fluxos apresentados neste capítulo, o P-MIA (Padrão Múltiplas Instâncias Autoadaptáveis), desenvolvido nesta Tese, atende às necessidades de áreas com as características já definidas. A Tabela 2 apresenta os padrões de dados, identificados como P e seu respectivo número, por exemplo, P4 (refere-se ao padrão de dados de número 4), comparados ao P-MIA.

Tabela 2 – Características de um Padrão de Dados para Autoadaptação de *Workflows* Científicos

Requisitos	P4	P8	P14	P15	P16	P27/28	P35	P37	P40	P-MIA
Manipulação de múltiplas instâncias	V	-	-	-	-	-	-	-	-	V
Visualização dos dados fora do ambiente de *workflow*	-	V	-	-	-	-	-	-	-	V
Troca de informações entre diferentes casos de *workflow*	-	-	V	-	-	-	-	-	-	V
Passagem de dados para um recurso ou serviço no ambiente operacional	-	-	-	V	-	-	-	-	-	V
Passagem de dados do ambiente operacional para o sistema de *workflow*	-	-	-	-	V	-	-	-	-	V
Transferência de dados por valor de um componente do *workflow* para outro	-	-	-	-	-	V	-	-	-	V
Interpretação de uma pré-condição para a execução de uma tarefa	-	-	-	-	-	-	V	-	-	V
Definição de uma pós-condição para a continuidade do processo	-	-	-	-	-	-	-	V	-	V
Definição de fluxo com base em condições	-	-	-	-	-	-	-	-	V	V
Manipulação de agrupamentos de dados	-	-	-	-	-	-	-	-	-	V
Determinação da instanciação de um processo com base em resultados já obtidos	-	-	-	-	-	-	-	-	-	V
Determinação da continuidade de um processo com base em resultados já obtidos	-	-	-	-	-	-	-	-	-	V
Determinação da alteração de prioridade de uma execução com base em resultados já obtidos	-	-	-	-	-	-	-	-	-	V

Assim, esse estudo serviu como subsídio para a definição do P-MIA a exemplo do padrão definido por Nardi em [NAR09]. Diferencia-se o padrão apresentado nesta Tese do definido por Nardi pela utilização de padrões de dados e não padrões de controle de fluxo, pois o problema (de Bioinformática) estudado manipula grandes volumes de dados e, além disso, exige a adaptação do *workflow* em tempo de execução, a qual é realizada com base na análise de valores obtidos a partir do processamento desses dados, na medida em que forem sendo produzidos. É importante destacar que o padrão definido por Nardi pode ser utilizado como base para o processamento em paralelo do padrão desenvolvido nesta Tese. O Capítulo 5 apresenta a formalização do P-MIA. O funcionamento do padrão e os testes realizados são detalhados nos capítulos 6 e 7 respectivamente.

5 FORMALIZAÇÃO DO P-MIA: PADRÃO MÚLTIPLAS INSTÂNCIAS AUTOADAPTÁVEIS

Este capítulo contém a formalização do Padrão Múltiplas Instâncias Autoadaptáveis, sua modelagem gráfica por meio de redes de Petri coloridas, os padrões de dados utilizados como base para sua definição e como é realizada a manipulação dos dados pelo padrão.

5.1 Componentes do Padrão

Definição 1: (*Snapshot* e Conjunto de *Snapshots*)

Seja S um conjunto não-vazio e finito de *snapshots*, produzidos por um simulador de dinâmica molecular para uma molécula de interesse (no caso, InhA). Seja $s \in S$ um *snapshot* e #S o número de *snapshots* de S.

Seja $F : S \rightarrow n$, uma função de mapeamento que mapeia cada *snapshot* s em um conjunto C_i.

Seja $C_i \subseteq S$, $1 \leq i \leq n$, $n \leq \#S$, um subconjunto de *snapshots* que satisfaz um critério de similaridade, definido pela função F(), onde n é o número de subconjuntos de S. F() é uma função de agrupamento que mapeia cada *snapshot* em um dos n conjuntos de *snapshots*.

Define-se, portanto:

- $\forall\ s \in S, \exists\ C_i, 1 \leq i \leq n \mid s \in C_i$.
- $\forall i,j, 1 \leq i \leq n \wedge 1 \leq j \leq n, i \neq j, C_i \cap C_j = \varnothing$.
- $C_1 \cup \ldots \cup C_n = S$.
- $C_i \neq \varnothing$.

*Comentário***:** A Figura 50 esquematiza a aplicação da função de similaridade sobre os *snapshots* e a definição dos subconjuntos, a partir daqui denominados de grupos.

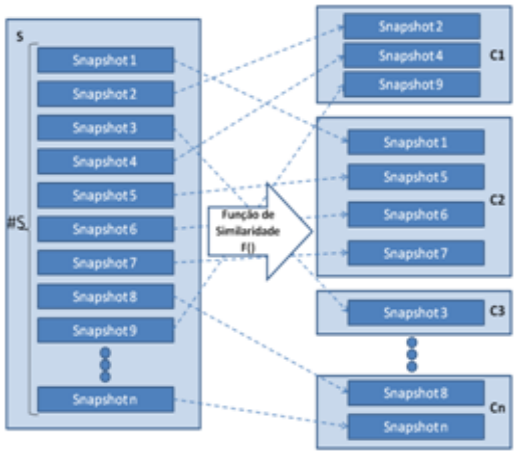

Figura 50 – Divisão de *Snapshots* em Grupos

Na Figura 50 verifica-se a existência de um conjunto de *snapshots*, definido por S, a quantidade total desses *snapshots* é dada por #S. A função de similaridade, responsável pela separação dos grupos gerou C1, C2, C3 e Cn. Nota-se que a quantidade de *snapshots* de cada grupo não é fixa, e que cada *snapshot* é mapeado para apenas 1 agrupamento.

Definição 2: (Execução de Programa de *Workflow* Científico)

Um programa de *Workflow* Científico é um programa de computador P que tem como entrada um conjunto de dados P.E e, como saída um conjunto de dados P.S.

Dadas 2 execuções distintas de P, P1 e P2:

- $P1.E = P2.E \rightarrow P1.S = P2.S$

Comentário: Execuções de Programas de *Workflows* Científicos precisam ser repetíveis. P.S também é chamado de resultado da execução.

Definição 3: (Lote de Execução de Programa de *Workflow* Científico)

Um lote L de execuções de P, programa de *workflow* científico, é um conjunto de m execuções independentes de P, podendo ser disparadas em conjunto, onde:

- $\forall i,j,\ 1 \leq i \leq m \wedge\ 1 \leq j \leq m,\ m \leq \#S,\ i \neq j,\ P_i.E \neq P_j.E$
- L.m é o número de execuções de L em P

L_{ix} é um dos x lotes de execução de um grupo, onde:

- x é a quantidade de lotes de C_i, $1 \leq i \leq n$, $n \leq \#S$
- $1 \leq x \leq \#C_i$

Comentário: As execuções de P, para um lote L, usam dados diferentes e podem ser executadas em paralelo, caso haja recursos computacionais adequados. A quantidade total de lotes é

determinada em tempo de execução, seguindo critérios pré-definidos como a quantidade mínima de *snapshots* a serem processados e o percentual de *snapshots* (amostragem) a serem processados de um grupo.

Definição 4: (Lote Residual)

Seja $\#L_i$ o número de *snapshots* de todos os lotes já definidos para C_i, $1 \leq i \leq n$, $n \leq \#S$. Seja $\#C_i$ o número de *snapshot* de C_i. Um lote Lr_i de execuções de P, programa de *workflow* científico, é um lote residual de um grupo C_i se:

- $\#C_i - \#L_i = Lr_i \leftrightarrow \#C_i - \#L_i <> \varnothing$

Comentário: Lotes residuais são formados toda vez que for criado um lote a partir de um conjunto de *snapshots*, desde que restem *snapshots* que não façam parte dos lotes já criados. Sobre esse lote residual aplicam-se regras de criação de novos lotes para o mesmo grupo, até que os critérios estabelecidos de quantidade mínima de *snapshots* e amostragem sejam atingidos.

Definição 5: (Quantidade Mínima de *Snapshots* e Amostragem)

Seja L_{ij}, $1 \leq i \leq n$, $n \leq \#S$, $1 \leq j \leq m$, $m \leq \#C_i$, um lote de execuções de P, correspondendo ao i-ésimo subconjunto de *snapshots* C_i.

Seja s_k, $1 \leq k \leq n$, $n \leq \#S$, um *snapshot* individual.

Seja ζ o número mínimo de execuções de P, para um lote L_{ij} qualquer.

Seja σ a quantidade, em percentual, de *snapshots* de um C_i ou de um Lr_i, a serem envolvidos em execuções de P em um lote L_{ij} qualquer. Portanto:

- Para cada C_i, criar um conjunto de lotes de execução L_{ij}, tal que:
  - Em cada execução de P, $P.E = s_k \in C_i$
  - $\#L_{ij} \geq \zeta$
  - $\#L_{ij} \geq \sigma$

Comentário: Os valores para quantidade mínima e amostragem são determinados antes do início da subdivisão dos grupos em lotes. Esses valores são utilizados para definir quantos *snapshots* compõem cada um dos lotes de um grupo.

Definição 6: (Status de *Snapshot* e Transição Válida de Estados de *Snapshots*)

Seja St um atributo de s, onde St = {A – ativo; F – Processado; D – Descartado; P – Prioridade Reduzida}

Portanto:

- $s.St = A \vee s.St = F \vee s.St = D \vee s.St = P$
- Se $(s.St = A) \rightarrow s.St = F$
- Se $(s.St = A) \rightarrow s.St = D$
- Se $(s.St = A) \rightarrow s.St = P$
- Se $(s.St = P) \rightarrow s.St = A$

Comentário: Atributos de status dos *snapshots*, cujos valores sejam iguais a D ou F, significam que os *snapshots* não serão mais executados pelo Programa de *Workflow* Científico. A transição entre possíveis valores do atributo *status* é encontrada na Figura 51, contendo o diagrama de transição de estados.

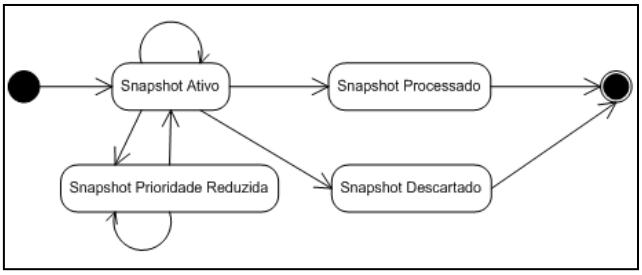

Figura 51 – Diagrama de transição de estados, identificando os status possíveis para processamento dos *snapshots*

Definição 7: (Arquivo ou Tabela de Acompanhamento do Processamento dos *Snapshots*)

Seja At um arquivo ou tabela que contenha todas as informações dos *snapshots*. Seja #s o número que corresponda à identificação de um s. Portanto:

- $<\#s, C_i, L_{ij}, St> \in At, 1 \leq i \leq n, n \leq \#S$

*Comentário***:** A definição de utilização de arquivo ou tabela para o armazenamento das informações depende da forma de implementação do modelo. Entretanto, as informações são fundamentais para o funcionamento e análise dos resultados.

Definição 8: (Melhor Valor e Pior Valor)

Seja $P.S.\sigma_i$, $1 \leq i \leq n$, $n \leq m$, onde m é uma amostra de C, o resultado do processamento de cada *snapshot* em um sistema de *workflow* científico. Seja MP.S o maior valor entre todos os P.S gerados pela execução dos *snapshots* da amostra. Seja PP.S o pior valor entre todos os P.S gerados pela execução dos *snapshots* da amostra. Portanto:

- $\exists\ MP.S \subseteq P.S.\sigma_i \mid \forall\ P.S.\sigma \subseteq P.S; MP.S \geq P.S.\sigma$

- $\exists$ PP.S $\subseteq$ P.S.σ_i | $\forall$ P.S.$\sigma \subseteq$ P.S; PP.S $\leq$ P.S.σ

Definição 9: (Análise Horizontal e Vertical, e Análise dos Resultados)

Seja AH, análise horizontal, a comparação dos diferentes resultados (P.S) dos lotes de um grupo ($C_i \rightarrow L_{ij}$, $1 \leq i \leq n$, $n \leq \#S$, $1 \leq j \leq m$, $m \leq \#C_i$). Seja AV, análise vertical, a comparação dos diferentes resultados (P.S) dos diferentes grupos de S ($C_i \subseteq S$, $1 \leq i \leq n$, $n \leq m$).

Seja N o número de *snapshots* (#s) já processados de um lote (L_{ij}). Seja P.S o resultado do processamento de um *snapshot*, após a submissão em um programa de *workflow* científico. Seja ΣP.S.L_{ij}, $1 \leq i \leq n$, $n \leq \#S$, $1 \leq j \leq m$, $m \leq \#C_i$, o somatório dos resultados dos *snapshots* de L_{ij} após serem submetidos à execução em um programa de *workflow* científico. Seja $\#L_{ij}$ a quantidade de *snapshots* de um lote (L_{ij}). Seja MP.S o melhor valor utilizado como parâmetro para análise dos resultados, gerado por amostragem. Seja PP.S o pior valor utilizado como parâmetro para análise dos resultados, gerado por amostragem.

Seja x a média aritmética dos resultados (P.S) dos *snapshots* já processados. Portanto:

- $x = \Sigma P.S.L_i / N$

Seja xMP a média aritmética dos resultados (P.S), entre MP.S e PP.S. Portanto:

- $xMP = \Sigma(MP.S, PP.S) / 2$

Seja xE a média aritmética estimada dos resultados já obtidos com o processamento dos *snapshots* do lote, no programa de *workflow* científico (P.S), e com os prováveis resultados dos demais *snapshots* do lote, cuja quantidade é chamada de f ($N - \#L_{ij} = f$). Para obter xE utiliza-se o valor do desvio padrão (α), dos diferentes P.S já obtidos, por meio da Regra Empírica adaptada, superestimando os valores de forma que se obtenha um resultado próximo do satisfatório. Portanto:

- $\forall$ xE, $\exists$ xE68 | xE68 = (68%f.$(x-2\alpha)$)
- $\forall$ xE, $\exists$ xE95 | xE95 = (95%f.$(x-3\alpha)$)
- $\forall$ xE, $\exists$ xE99,7 | xE99,7 = (99,7%f.$(x-4\alpha)$)
- $xE = (\Sigma P.S.L_i + xE68 + xE95 + xE99,7)/\#L_{ij}$

A definição de status dos *snapshots*, a partir da análise dos resultados é feita da seguinte forma:

- $\forall$ P.S | $((P.S > x) \wedge (P.S > xE)) \rightarrow St = A$
- $\forall$ P.S | $((P.S > x) \wedge (P.S <= xE)) \rightarrow St = P$

- $\forall$ P.S | ((P.S < x) ^ (P.S <= xE)) $\rightarrow$ St = D

Comentário: (i) a interseção de *snapshots* que pertencem a xE68 e xE95 e a xE95 e xE99,7, deve ser eliminada para o cálculo da expressão; (ii) a análise dos resultados fará com que se tenha alteração no status (St) dos *snapshots* (s); (iii) a regra empírica adaptada é apresentada no Capítulo 6.

Definição 10: (Arquivo ou Tabela com Resultados do Processamento dos Lotes)

Seja Atp um arquivo ou tabela que contenha todas as informações dos processamentos dos lotes. Seja C_i, o número que corresponda à identificação de um C. Seja x_i a média aritmética de cada grupo, considerando os *snapshots* processados até o momento da análise. Seja xE_i a média aritmética estimada de cada grupo, considerando os *snapshots* processados até o momento da análise. Seja St_i o status de cada grupo, considerando os *snapshots* processados até o momento da análise. Portanto:

- $<C_i, x_i, xE_i, St_i> \in$ Atp, $1 \leq i \leq n$, $n \leq \#S$

*Comentário***:** A definição de utilização de arquivo ou tabela para o armazenamento das informações depende da forma de implementação do modelo. Entretanto, as informações são fundamentais para o funcionamento e análise dos resultados.

Definição 11: (P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis)

Formalmente, o P-MIA é uma tupla P_MIA = <C, L, s, P.S, MP.S, PP.S>, onde:

- C = $\{C_1, C_2, ..., C_m\}$ conjunto finito de grupos de *snapshots*;
- L = $\{L_1, L_2, ..., L_n\}$ conjunto finito de lotes formados a partir de um grupo individual de *snapshots*;
- s é o *snapshot* contido em um lote de um grupo
  - $s \subseteq L \mid L \subseteq C$
- P.S = $\{P.S_1, P.S_2, ..., P.S_m\}$ conjunto finito de resultados do processamento de cada um dos *snapshots*;
- MP.S é o melhor valor dentre todos os valores processados de uma amostra de *snapshots*:
  - $MP.S \subseteq P.S$
- PP.S é o pior valor dentre todos os valores processados de uma amostra de *snapshots*:

- PP.S $\subseteq$ P.S

- MP.S $\neq$ PP.S

Comentário: A tupla é composta pelos principais componentes do padrão. Características importantes como as funções utilizadas para definição de prioridade não são representadas na tupla explicitamente, pois fazem parte do funcionamento.

5.2 P-MIA Modelado com Redes de Petri Coloridas

Com o objetivo de simular o funcionamento do modelo e de se ter uma representação gráfica, o P-MIA foi modelado por meio de redes de Petri Coloridas, utilizando para isso a ferramenta CPN Tools [CPN10], com a qual é possível editar, analisar e simular Redes de Petri coloridas. Com essa ferramenta pode-se fazer várias implementações de sistemas, desde uma rede até um sistema completo. Um modelo inicial, representando o P-MIA em alto nível pode ser visualizado na Figura 52. Nessa figura, são representados, apenas, os grupos, a separação desses grupos em lotes e a execução dos *snapshots*, sem a preocupação com as análises intermediárias do modelo. Na Figura 52, existe apenas um conjunto de dados 1`(1,2,3,4). Nesse conjunto de dados o primeiro valor está representando a quantidade de grupos, ou seja, 1; o segundo valor está representando a quantidade de lotes, ou seja, 2; o terceiro e quarto valores são os lotes em si, cujos valores 3 e 4 referem-se à quantidade de *snapshots* de cada um dos dois lotes.

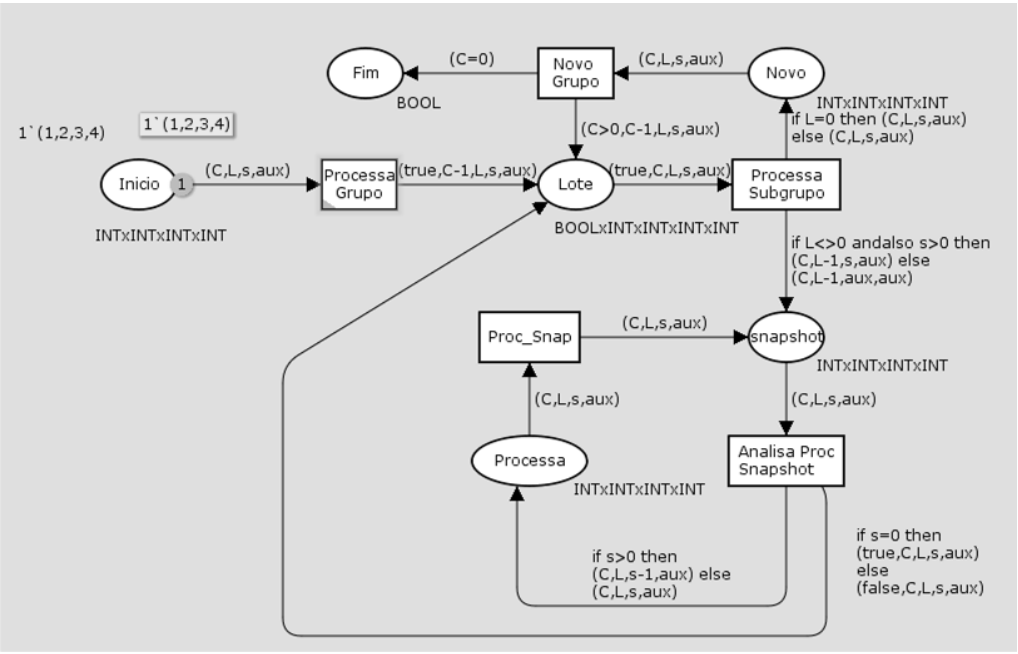

Figura 52 – P-MIA em alto nível modelado com Redes de Petri Coloridas, utilizando a ferramenta CPN Tools

Com o objetivo de melhor representar o funcionamento do P-MIA, com as análises intermediárias, a Figura 53 apresenta a modelagem em redes de Petri com maior nível de detalhamento, sem considerar dados externos como arquivos ou tabelas em bancos de dados. Nessa representação, são feitas as análises sobre os resultados obtidos, identificando as diferentes possibilidades: Altera Prioridade, Finaliza Lote, Descarta Lote ou Mantém Prioridade. A transição denominada *"Armazena resultado – gera média"*, da Figura 53 é a responsável por fazer uma das comunicações com o ambiente externo, buscando informações de resultados já armazenados. A partir do resultado (saída) obtido com esse processamento, as demais transições serão executadas. Na Figura 53 se tem o conjunto de dados 1`(1,1,5), correspondendo a 1 grupo, 1 lote e 5 *snapshots* dentro desse lote. A modelagem apresentada nessa figura também foi feita com a ferramenta CPN Tools [CPN10].

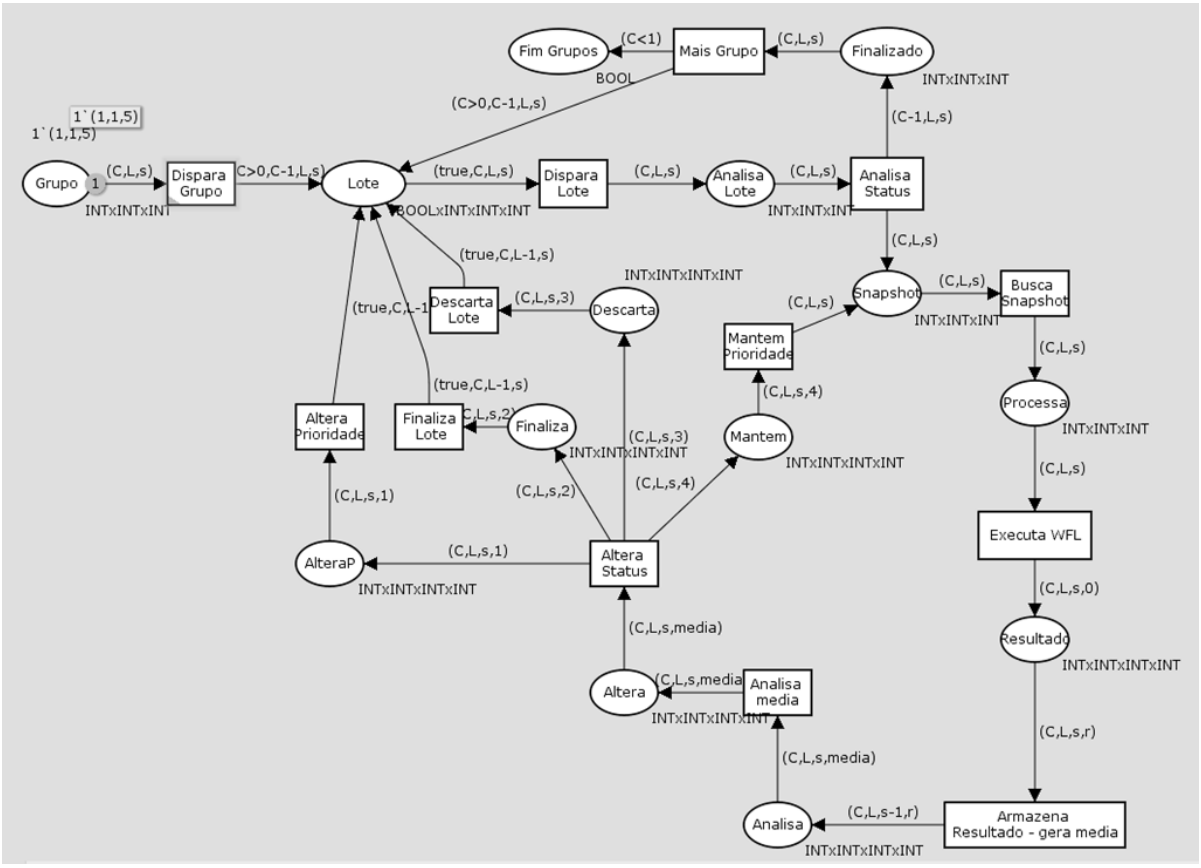

Figura 53 – P-MIA detalhamento da análise do resultado modelado com Redes de Petri Coloridas, utilizando a ferramenta CPN Tools

5.3 Padrões de Dados no P-MIA

O P-MIA utiliza, como base para sua elaboração, 10 padrões de dados definidos por Russel et al. em [RUS04, RUS05]. Cada padrão, bem como sua utilização, é detalhado nas próximas seções.

5.3.1 Padrão 4. Dados de Múltiplas Instâncias

Considerando que uma das características desse padrão, já apresentada anteriormente, é que uma tarefa seria designada como sendo tarefa de múltiplas instâncias quando, uma vez habilitada, múltiplas instâncias poderiam ser iniciadas simultaneamente, o P-MIA utiliza esse padrão, uma vez que considera a possibilidade de execuções de suas tarefas em paralelo. Assim, várias tarefas estariam em execução, com diferentes instâncias, ao mesmo tempo.

5.3.2 Padrão 8. Dados de Ambiente

Esse padrão considera que dados estejam em um ambiente externo, acessados pelos componentes do *workflow* durante a execução. O P-MIA manipula a característica desse padrão quando considera o acesso a arquivos armazenados em estruturas de diretórios no sistema operacional ou a tabelas organizadas em bancos de dados. A principal utilização desse padrão está no armazenamento dos resultados obtidos pelos *snapshots* para posterior análise das médias geradas, com o objetivo de definir as próximas etapas de execução dos *snapshots* e a alteração ou não de seus status.

5.3.3 Padrão 14. Interação de Dados – Casos para Casos

Esse padrão é utilizado pelo P-MIA ao considerar a possibilidade de acesso, por diferentes casos de *workflow* (instâncias em execução), de uma mesma informação por meio de uma estrutura de armazenamento compartilhada. Dessa forma, diferentes casos em execução do P-MIA podem estar utilizando dados armazenados externamente ao programa de execução de *workflow* científico para realizarem suas análises.

5.3.4 Padrão 15. Interação de Dados – Tarefas para Ambiente Externo – *Push-Oriented*

Esse padrão prevê a habilidade de uma tarefa iniciar a passagem de dados para um recurso ou serviço no ambiente operacional. Conforme definido pelo padrão existem duas categorias principais que subsidiam a implementação desse tipo de interação: (i) mecanismo de integração explícito: onde o sistema de *workflow* provê construtores específicos para passagem de dados ao ambiente externo; (ii) mecanismo de integração implícito: onde a passagem dos dados ocorre implicitamente dentro de implementações que fazem chamadas nos processos do *workflow* e não são suportadas diretamente pelo ambiente. O P-MIA utiliza-se do mecanismo de integração explícito, pois se sabe exatamente quais são os momentos em que é necessária a passagem de dados para o ambiente externo. Esses momentos são: obtenção do resultado do processamento de cada *snapshot*; alteração dos status dos *snapshots* e dos lotes de processamento.

5.3.5 Padrão 16. Interação de Dados – Ambiente Externo para Tarefas – *Pull-Oriented*

Nesse padrão, uma tarefa de *workflow* requisita dados de recursos ou serviços de um ambiente operacional. Pode envolver o acesso aos dados de um repositório, por exemplo, e é principalmente essa a característica utilizada pelo P-MIA, quando dados armazenados externamente (resultados dos processamentos dos *snapshots*) necessitam ser acessados para que se realizem as médias e se prossiga com as análises.

5.3.6 Padrão 27. Transferência de Dados por Valor – Entrada

A habilidade de um componente do *workflow* receber dados de entrada por valor de outro componente é outra característica do P-MIA. Quando uma tarefa obtém um resultado de saída e esse resultado deve ser encaminhado como entrada para outra tarefa, esse padrão é utilizado.

5.3.7 Padrão 28. Transferência de Dados por Valor – Saída

A habilidade de um componente do *workflow* passar dados por valor para outro componente é outra característica do P-MIA, pois segue o mesmo princípio da transferência de dados por valor – entrada. Quando uma tarefa obtém um resultado de saída e esse resultado deve ser encaminhado como entrada para outra tarefa, esse padrão é utilizado.

5.3.8 Padrão 35. Pré-condição para Tarefa – Valor de Dados

Pré-condições com base em dados podem ser especificadas por tarefas que definem sua execução pela presença de um determinado valor para o dado, especificado em tempo de execução. Esse padrão também é muito utilizado pelo P-MIA, pois a alteração dos status dos *snapshots* e a execução das diferentes ações: alteração de prioridades, descarte dos *snapshots* etc. são definidas a partir do valor de dados.

5.3.9 Padrão 37. Pós-Condição para Tarefa – Valor de Dados

Pós-condições com base em dados podem ser especificadas por tarefas que definem sua execução por valores de parâmetros específicos em tempo de execução. Da mesma forma que o padrão 35, esse padrão também é utilizado pelo P-MIA.

5.3.10 Padrão 40. Roteamento Baseado em Dados

A habilidade de alterar o fluxo dentro de um caso de *workflow* após a análise de expressões baseadas em dados. Esse padrão serve como uma agregação dos dois maiores padrões de controle de fluxo que dependem de dados: (i) escolha exclusiva, onde o fluxo de controle é passado para uma das muitas tarefas seguintes, dependendo da saída de uma decisão ou do valor

de uma expressão; (ii) escolha múltipla, onde dependendo da saída de uma decisão ou do valor de uma expressão, o fluxo de controle é passado para várias tarefas seguintes. O P-MIA utiliza-se principalmente da escolha exclusiva, pois a análise dos resultados obtidos pode direcionar o fluxo a diferentes caminhos.

5.3.11 Características Específicas do P-MIA

Além dos padrões de dados definidos por Russel et al. [RUS04, RUS05], o P-MIA também possui outras características. São elas:

- manipulação de agrupamentos de dados;
- determinação da instanciação de um processo com base em resultados já obtidos;
- determinação da continuidade de um processo com base em resultados já obtidos;
- determinação da alteração da prioridade de execução com base em resultados já obtidos.

5.3.11.1 P-MIA 1: Manipulação de Agrupamentos de Dados

Essa característica é utilizada toda vez que for necessário o processamento de um grande volume de dados, agrupados conforme um critério de similaridade qualquer. Possui duas possibilidades de funcionamento: (i) todo o grupo é submetido ao processamento; (ii) é realizada a divisão do grupo, em partes menores, antes da submissão ao processamento. A Figura 54 apresenta a manipulação de grupos completos de dados e a Figura 55 apresenta a manipulação de lotes menores criados a partir dos grupos. Tanto a representação da Figura 54, quanto à da Figura 55 foram modeladas com redes de Petri coloridas, utilizando a ferramenta CPN Tools [CPN10].

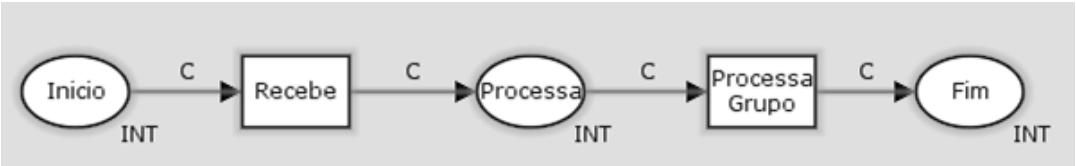

Figura 54 – P-MIA: manipulação de agrupamento de dados – completo

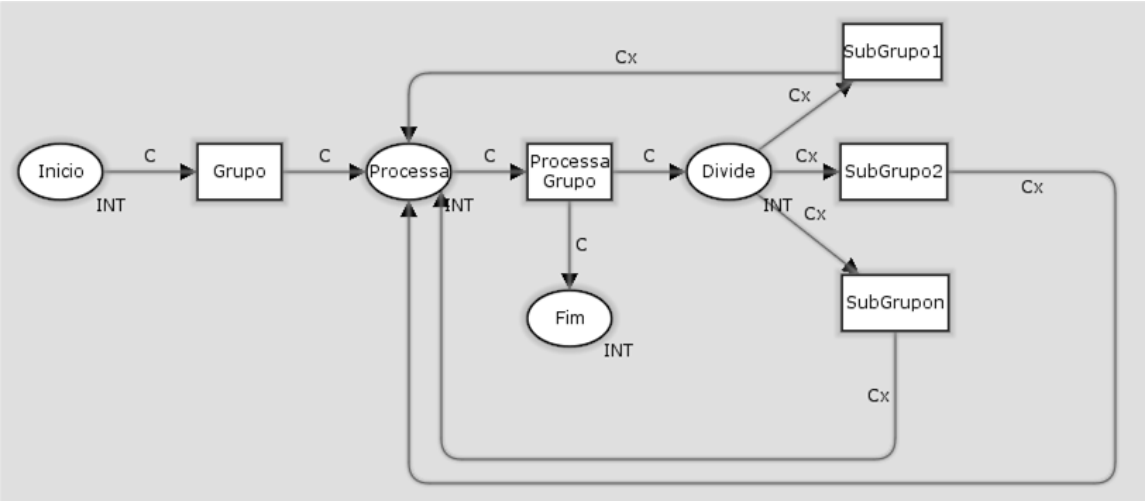

Figura 55 – P-MIA: manipulação de agrupamento de dados – em partes menores

5.3.11.2 P-MIA 2: Determinação da Instanciação de um Processo com Base em Resultados já Obtidos

Essa característica, muito semelhante em sua essência ao padrão 40, definido por Russel et al. [RUS04, RUS05], diferencia-se, principalmente, em sua aplicação. Os testes, para a definição pela instanciação de um processo ou não, são realizados antes de qualquer execução de tarefa do *workflow*. Portanto, pode-se definir que as tarefas do *workflow* estariam encapsuladas e que sua execução seria determinada por uma ação externa. A Figura 56 esquematiza o funcionamento dessa característica, modelado por meio de redes de Petri coloridas.

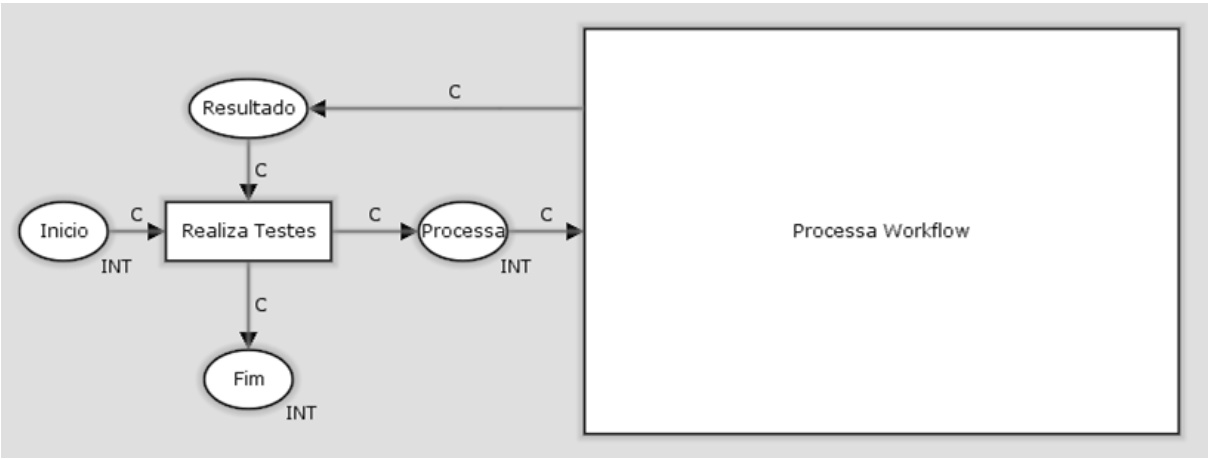

Figura 56 – P-MIA: determinação da instanciação de um processo com base em resultados já obtidos

5.3.11.3 P-MIA 3: Determinação da continuidade de um processo com base em resultados já obtidos

Essa característica, também é muito semelhante ao padrão 40, definido por Russel et al. [RUS04, RUS05]. A diferença está na aplicação direcionada, pois a determinação da continuidade de um processo está na determinação de continuidade de execução dos dados de um

determinado lote. A análise dos resultados e dos status dos dados de um lote é fundamental para se inferir sobre a continuidade de um processo. Sua representação é a mesma de escolha exclusiva.

5.3.11.4 P-MIA 4: Determinação da alteração da prioridade de execução com base em resultados já obtidos

Essa característica utiliza-se da manipulação de informações que estão no meio externo. Seu ponto principal está na alteração ou não de atributos dos dados envolvidos no processo. Não existe, nesse caso, exclusão ou descarte de dados a serem processados. O que se pode ter é um atraso na execução. Os dados a serem alterados são definidos em tempo de execução, com base no processamento de resultados já obtidos.

5.4 Considerações do Capítulo

Este capítulo apresentou a formalização do padrão desenvolvido nesta Tese: P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis. Para a definição do padrão, utilizou-se como base termos da formalização tradicional de redes de Petri, bem como a representação gráfica do P-MIA em redes de Petri coloridas, desenvolvida por meio da ferramenta CPN-Tools [CPN10]. A vantagem em se formalizar o padrão está na validação do modelo, identificando seus principais componentes e na utilização de uma linguagem com grande poder de expressão [GUB06]. Na busca por facilitar a implementação do modelo, o Capítulo 6 oferece um detalhamento do funcionamento do padrão, exemplificando e apresentando as funções definidas, em especial, para a análise dos resultados.

Este capítulo também apresentou os padrões de dados, dos diferentes apresentados no Capítulo 4, que subsidiaram o modelo aqui proposto, bem como o detalhamento de características específicas do P-MIA que não são encontradas em outros padrões já definidos na literatura.

6 P-MIA: PADRÃO MÚLTIPLAS INSTÂNCIAS AUTOADAPTÁVEIS - UM PADRÃO DE DADOS PARA AUTOADAPTAÇÃO DE INSTÂNCIAS EM EXECUÇÃO EM WORKFLOWS CIENTÍFICOS

No Capítulo 3 foi apresentada a área de aplicação, na qual foram realizados os experimentos, apresentados no Capítulo 7, sobre o padrão desenvolvido nesta Tese. No Capítulo 3 também foi apresentada a atuação do LABIO e destaca-se, neste momento, o trabalho desenvolvido por Karina Machado em [MAC07a], de um *workflow* científico para a modelagem do processo de desenvolvimento de fármacos assistido por computador, utilizando receptor flexível. O funcionamento completo do *workflow* científico, desenvolvido por Karina Machado pode ser encontrado em [MAC07a]. O fluxo pode ser visualizado na Figura 57.

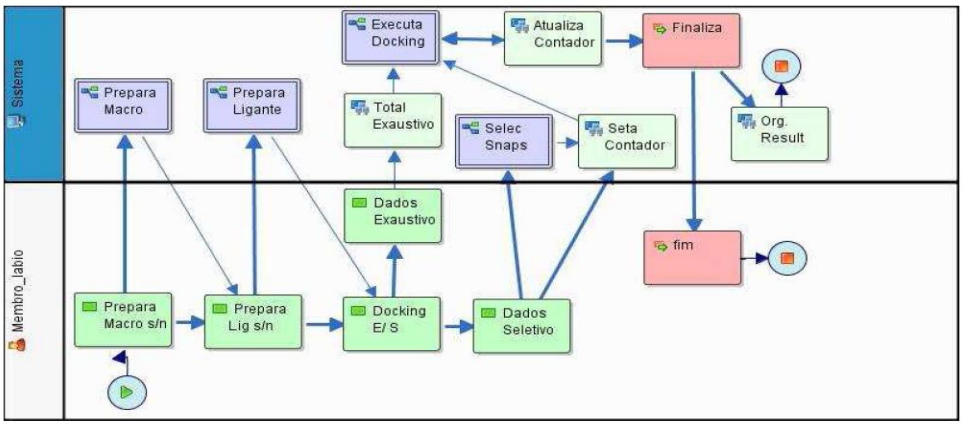

Figura 57 – Modelagem final do processo de desenvolvimento de fármacos assistido por computador, desenvolvido por Karina Machado em [MAC07]

O *workflow* científico desenvolvido por Machado [MAC07a] é apresentado em maiores detalhes nesta Tese por ser motivador de seu desenvolvimento. Considera-se o desenvolvimento do *workflow* científico como um ganho considerável de processamento, pois automatizou etapas antes manuais, envolvendo a execução de diferentes sistemas. Apesar disso, existem algumas etapas, no *workflow* desenvolvido por Machado, que demandam maior atenção. Conforme a própria autora:

> "Apesar dos experimentos realizados mostrarem que a seleção pela energia se mostra eficiente, utilizando-se ligantes de uma mesma classe, ainda são muitos *snapshots* que precisam ser utilizados (foram utilizados 1.000 *snapshots* durante os experimentos seletivos). Se utilizarmos tantos *snapshots* em experimentos com milhares de ligantes, o tempo necessário para terminar a execução dos experimentos é muito grande (se para um ligante o tempo para executar o docking seletivo é em torno de 100 hs, se forem utilizados 100 ligantes, já são necessários 10.000 hs ou 417 dias de processamento em um PC com uma configuração semelhante aos PC LABIO 1 e PC LABIO 2 utilizados nos nossos experimentos)." [MAC07]

Karina refere-se às etapas envolvidas em torno da etapa *Dados Seletivo*, visualizada na Figura 57, que inferiu ganho de processamento ao *workflow*, antes executado apenas de forma exaustiva. Nessa etapa, a autora utiliza uma técnica de seleção de *snapshots* baseada na energia livre de ligação (FEB), podendo ser utilizada, apenas, após um experimento exaustivo, selecionando aqueles *snapshots* da macromolécula que resultaram nos melhores resultados de docagem com determinado ligante. Esses *snapshots* são utilizados para a docagem de outros ligantes pertencentes à mesma classe que o primeiro. Karina também afirma que há uma boa possibilidade de se encontrar bons resultados sem a necessidade de executar o experimento, considerando todos os *snapshots* da trajetória de DM da proteína. Por consequência, define-se que bons resultados obtidos por uma determinada amostragem de *snapshots* de um grupo, considerando agrupamentos por similaridade, podem ser utilizados para a determinação da autoadaptação dos *snapshots* no fluxo.

O trabalho desenvolvido nesta Tese busca contribuir com a redução da quantidade de experimentos a serem executados, via a definição de um padrão capaz de substituir as etapas que envolvem o processamento de dados seletivos no *workflow* desenvolvido por Karina Machado [MAC07a], de forma que não exista a necessidade de execuções exaustivas, tornando a seleção de *snapshots* uma etapa dinâmica. A substituição de algumas etapas do *workflow* desenvolvido por Karina pelo P-MIA é apresentada na Figura 58.

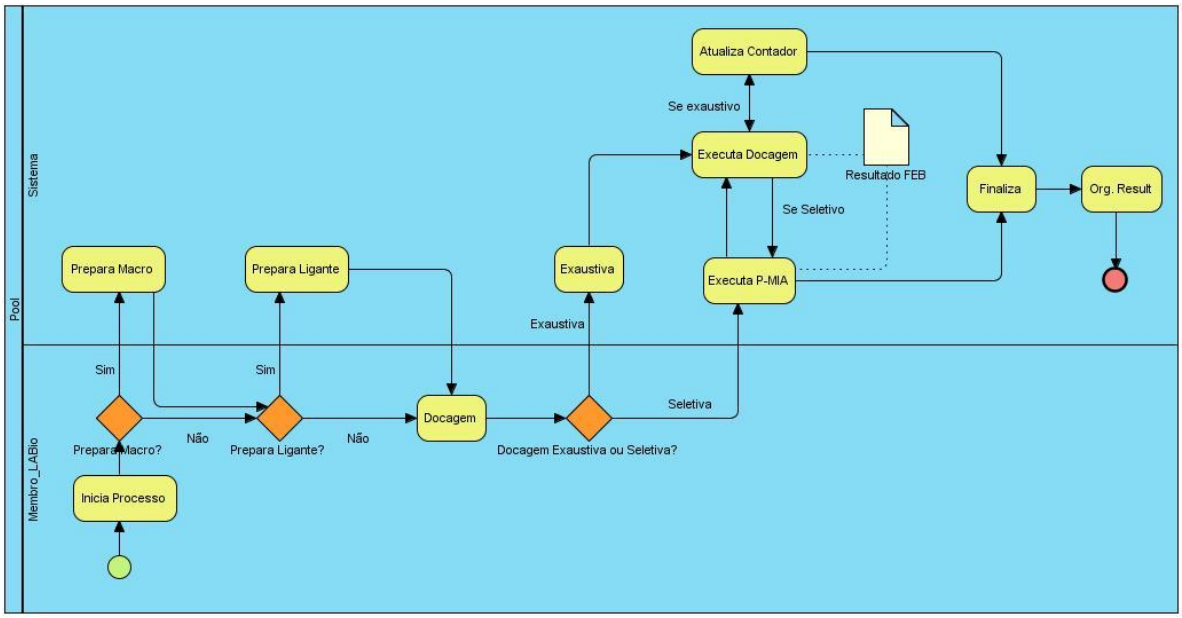

Figura 58 – Inserção do P-MIA no *workflow* científico desenvolvido por Karina Machado em [MAC07a], criado a partir da ferramenta Visual Architect, utilizando BPMN como notação

A Figura 58 contém o mapeamento de um *workflow* científico, com base no *workflow* de [MAC07a], modelado utilizando-se a notação denominada BPMN (*Business Process Modeling Notation*) [OMG09]. BPMN é uma notação visual para representação de fluxos de processos que pode ser mapeada para diversos formatos de execução. Proporciona às ferramentas o uso de uma representação gráfica padronizada, permitindo a divulgação dos processos de maneira uniforme, facilitando o entendimento entre profissionais e cientistas. Neste momento, por apresentar uma notação mais próxima da utilizada por Karina Machado e por representar claramente os pontos de decisão, foi a escolhida. A ferramenta utilizada para o mapeamento é denominada *Business Process Visual Architect Modeler Edition*[1].

Na Figura 58 a etapa *Executa P-MIA* substituiu as seguintes etapas da Figura 57: *Dados Seletivo, Selec Snaps, Seta Contador* e *Atualiza Contador* (quando utilizada para dados seletivos). O funcionamento do P-MIA é apresentado em maiores detalhes nas próximas seções deste capítulo.

O padrão apresentado nesta Tese, apesar de substituir etapas do *workflow* científico desenvolvido para a área de Bioinformática, pesquisada pelo LABIO, busca atender a outras áreas, desde que possuam as seguintes características:

- necessidade de manipulação de grandes volumes de dados;

[1] http://www.visual-paradigm.com/

- existência de dados com características semelhantes, podendo ser agrupados por alguma função de similaridade;
- existência de um resultado final após o processamento de cada um dos dados, em um sistema de *workflow* científico, representado por uma informação numérica, que permita comparar a qualidade das execuções entre si;
- existência de um parâmetro que possa ser considerado bom para a informação obtida com o processamento.

Para que o padrão obtenha os resultados esperados é fundamental a utilização de uma função de similaridade para a definição dos grupos de *snapshots*, que consiga agrupá-los considerando suas características, de forma que *snapshots* de um mesmo grupo tenham as mesmas chances de obterem resultados bons ou ruins.

Este capítulo, portanto, apresenta o funcionamento do Padrão Múltiplas Instâncias Autoadaptáveis (P-MIA), podendo ser implementado em aplicações de diferentes áreas, desde que possuam as características já definidas. O funcionamento do padrão é apresentado considerando sua utilização na área de Bioinformática, com o objetivo de direcionar o entendimento.

6.1 Conceitos Fundamentais

Para que se entenda o comportamento do P-MIA, é importante o detalhamento de algumas características utilizadas pelo padrão. A etapa preliminar de submissão dos dados ao Padrão Múltiplas Instâncias Autoadaptáveis prevê a separação de todos os *snapshots* envolvidos no processo, em grupos, por uma função de similaridade. A Figura 59 esquematiza a aplicação da função de similaridade sobre os *snapshots* e a definição dos grupos. É importante destacar que essa função de similaridade não faz parte do escopo deste trabalho, sendo apenas utilizada para a separação dos grupos, definidos como dados de entrada para o funcionamento do padrão proposto. Uma função de similaridade, direcionada à realidade da área de Bioinformática, está sendo desenvolvida por Karina Machado em sua Tese de Doutorado.

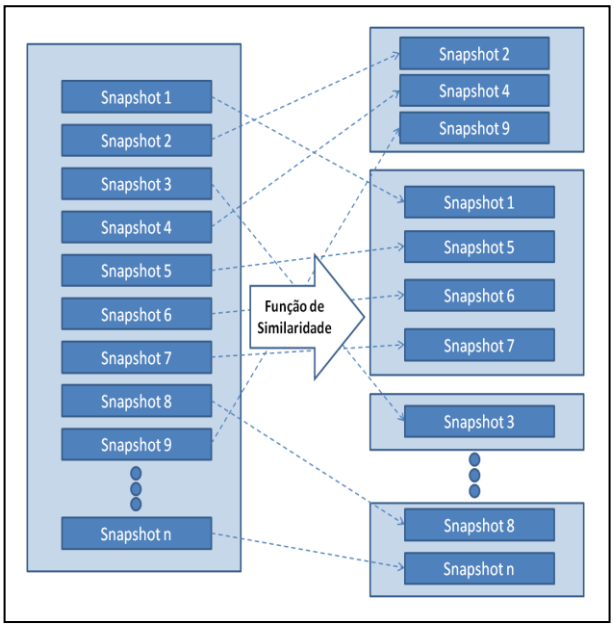

Figura 59 – Aplicação da Função de Similaridade

Na Figura 59 verifica-se a existência de um conjunto de *snapshots*, à esquerda da figura. A função de similaridade, responsável pela separação dos grupos é representada pela seta e, à direita da Figura 59, está representada a geração de quatro grupos, com quantidades diferentes de *snapshots*. Com a definição desses grupos, essas informações devem ser organizadas, podendo ser armazenadas sob a forma de um arquivo ou tabela, desde que contenha as seguintes informações:

- *Snapshot*: identifica o número do *snapshot* ao qual as demais informações se referem;
- *Grupo*: identifica o grupo ao qual o *snapshot* pertence;
- *Lote*: identifica o lote ao qual o *snapshot* pertence (melhor detalhado nas próximas seções);
- *Status*: identifica a situação sobre o processamento do *snapshot*, podendo conter um dos seguintes valores:
  - A – Ativo e aguardando pelo processamento
  - F – Processamento Finalizado
  - D – Descartado por resultado insatisfatório dos demais *snapshots* do grupo;
  - P – Prioridade do grupo ao qual o *snapshot* pertence foi alterada.

A informação referente ao status de cada *snapshot* é fundamental para a determinação de sua execução ou não. Considera-se, portanto, que apenas *snapshots* ativos são processados. As informações contidas no arquivo criado são utilizadas pelo padrão quando da análise dos resultados e da definição de prioridades e continuidade de processamento. A Tabela 3 exemplifica a estrutura do arquivo/tabela criado com as informações definidas anteriormente, após alguma execução. Nessa tabela se tem três *snapshots*, cada um de um grupo e de lotes diferentes. O *snapshot* de número 320 foi finalizado; o *snapshot* de número 1.457 foi descartado; e o *snapshot* de número 522 está ativo, aguardando a execução.

Tabela 3 – Estrutura do arquivo ou tabela com dados dos *snapshots* para acompanhamento e processamento

Snapshot	**Grupo**	**Lote**	**Status**
320	0	0	F
1457	1	2	D
522	0	1	A

O padrão também trabalha com os valores de variáveis, podendo ser definidos pelo usuário, antes do início de sua execução, e são representadas na Figura 64 pelo identificador “Referência”. Essas variáveis são:

- *Quantidade mínima* de *snapshots* a serem processados: o usuário fornece um valor numérico, correspondendo à quantidade mínima de *snapshots* que devem ser processados para a formação de cada lote de um grupo. Caso esse valor não seja fornecido, considera-se 50 como quantidade mínima para criação dos lotes, por ser possível criar grupos cujos resultados possam ser interpretados pelo padrão.
- *Amostragem*, em percentual, da quantidade de *snapshots* que formarão cada lote: o usuário fornece um valor numérico, correspondendo ao percentual do total (ou residual) de *snapshots* do grupo que formará cada lote. Caso esse valor não seja fornecido, considera-se 30% como percentual a ser utilizado para a criação dos lotes por ser possível criar grupos cujos resultados possam ser interpretados pelo padrão.

Os valores definidos previamente para as variáveis *quantidade mínima* e *amostragem* são utilizadas nos testes apresentados no Capítulo 7, comprovando sua viabilidade. O P-MIA prevê, ainda, a separação dos grupos definidos, por meio da função de similaridade, em lotes menores e

a análise dos resultados obtidos, definindo a alteração dos status dos *snapshots*. Essas características também são justificadas no Capítulo 7.

6.2 Separação em Lotes

Os *snapshots* que estão separados em diferentes grupos, cuja separação foi realizada pela aplicação de uma função de similaridade apropriada, devem ser subdivididos em lotes. Essa é uma característica importante do P-MIA. Esses lotes são utilizados como escopo para análises preliminares e intermediárias dos resultados obtidos. É importante destacar que se pode ter uma quantidade previamente desconhecida de lotes, pois essa quantidade é definida em tempo de execução, tendo como base os valores de referência definidos inicialmente e que correspondem à quantidade mínima e à amostragem. A adoção por lotes menores tem por base os testes encontrados no Capítulo 7, que concluiu que a análise de quantidades menores de dados fornece melhores resultados que a análise de um grupo como um todo. Além disso, leva-se em conta a *amostragem* e a *quantidade mínima* para que os lotes não sejam todos de um mesmo tamanho, conferindo maior flexibilidade ao modelo.

Quando da criação dos lotes deve-se ter atenção em relação ao número de *snapshots* que compõe cada um, considerando as seguintes etapas:

1. Um lote é formado pela quantidade de *snapshots* que correspondam ao percentual fornecido para amostragem; entretanto, caso essa quantidade seja menor que a mínima definida pelo usuário, o lote será criado com quantidade igual à mínima. Caso a quantidade total de *snapshots* do grupo seja menor que a mínima exigida, o lote será criado com a quantidade total de *snapshots*.
2. Os *snapshots* que não fizerem parte do primeiro lote formarão o que se chama de lote residual. Esse lote residual poderá ser submetido ao programa de *workflow* após análise dos resultados, retornando ao processamento definido no item (1).

Na Figura 60, por exemplo, pode-se observar a separação dos grupos de *snapshots*, C1, C2 e C3 apresentados na Figura 59, em lotes. Considera-se, para a interpretação da figura, que o percentual de amostragem definido pelo usuário seja 50% e que a quantidade mínima de *snapshots* a serem processadas por lote seja igual a 2.

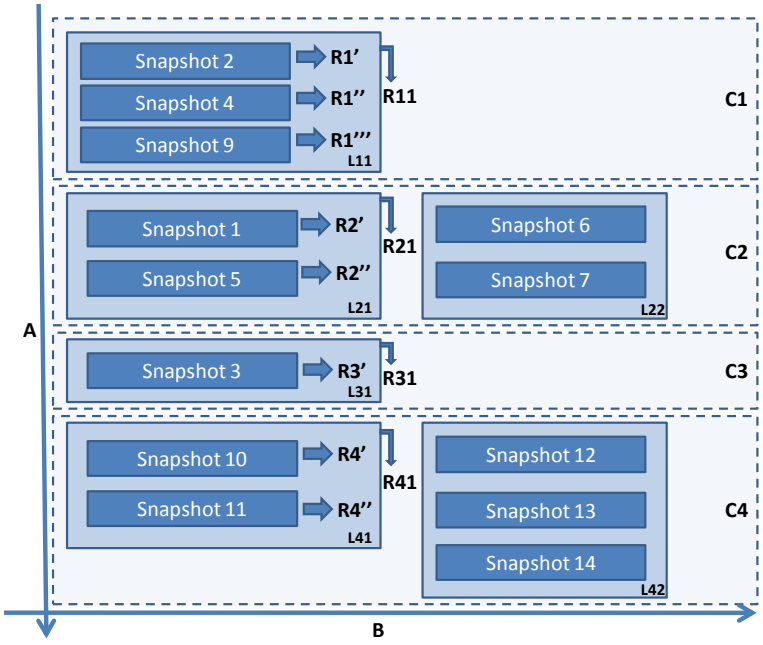

Figura 60 – Separação dos subgrupos em lotes e representação dos resultados individuais de cada *snapshot* e do lote como um todo

Na Figura 60, o grupo 1 (C1) possui três *snapshots*. O primeiro lote (L11), nesse caso, será composto pela quantidade total de *snapshots* do grupo, pois 50% sobre 3 é igual a 1,5, que é menor que a quantidade mínima definida (2). Se for considerado um lote com 2 *snapshots*, o lote residual ficaria com apenas 1 *snapshot*, também menor que a quantidade mínima. Para a criação dos lotes, em momento de execução, segue-se o algoritmo recursivo da Figura 61.

O algoritmo da Figura 61 deve ser executado para cada um dos grupos, gerados a partir de uma função de similaridade, e é importante destacar que a amostragem e a quantidade mínima de *snapshots* a serem executados permanecem constantes, não sendo alterados após a definição preliminar.

```
Definição_Lote (total_snapshost_grupo, amostragem, quantidade_mínima)
início
    Se ((amostragem * total_snapshots_grupo) < quantidade_mínima) então
       lote ← quantidade_mínima
    Senão lote ← amostragem;
    Se ((total_snapshosts_grupo - lote) < quantidade_mínima) então
       lote ← total_snapshosts_grupo
       .
       .
       .
    <conjuntos de comandos>
    .
    .
    .
    Se ((total_snapshosts_grupo - lote) >= quantidade_mínima) então
       Definição_lote(total_snapshosts_grupo - lote, amostragem, quantidade_mínima)
Fim.
```

Figura 61 – Algoritmo para definição dos lotes em momento de execução

As tabelas da Figura 62 demonstram três diferentes testes de mesa realizados com o algoritmo da Figura 61, com valores mais abrangentes, diferentes dos definidos para análise da Figura 60, considerando-se, sempre, a mesma quantidade total de *snapshots* e alterando, apenas, a amostragem e a quantidade mínima para processamento. Na Figura 62, os nomes das colunas das tabelas referem-se a:

- *Teste*: representa cada teste de mesa realizado;
- *Execução*: representa cada lote criado;
- *Amostragem*: percentual utilizado para a criação dos diferentes lotes de um grupo;
- *Quantidade_mínima*: quantidade mínima de *snapshots* utilizados para a criação dos diferentes lotes de um grupo;
- *Total_snapshots_grupo*: quantidade total de *snapshots* que compõem o grupo;
- *Lote*: quantidade de *snapshots* que formaram o lote, seguindo o algoritmo da Figura 61.

Teste	Execução	Amostragem	quantidade_mínima	total_snapshots_grupo	lote
1	1	40%	30	100	40
1	2	40%	30	60	30
1	3	40%	30	30	30

Teste	Execução	Amostragem	Quantidade_mínima	Total_snapshots_grupo	Lote
2	1	70%	30	100	70
2	2	70%	30	30	30

Teste	Execução	Amostragem	Quantidade_mínima	Total_snapshots_grupo	Lote
3	1	40%	40	100	40
3	2	40%	40	60	60

Figura 62 – Testes de mesa realizados sobre o algoritmo Definição_Lote

Os testes de mesa, da Figura 62, realizados a partir do algoritmo da Figura 61, podem ser detalhados da seguinte forma:

- No primeiro teste de mesa a quantidade total de *snapshots* do grupo é dividida em três lotes para execução. O primeiro lote possui a quantidade de *snapshots* definido pelo percentual de amostragem, pois 40% de 100 é igual a 40, maior que a quantidade mínima (30) definida. No segundo lote, a quantidade de *snapshots* é igual à quantidade mínima, pois 40% de 70

snapshots (lote residual) é igual a 28, menor que o mínimo definido, e esse é o mesmo princípio para a definição da quantidade do terceiro lote.

- No segundo teste a quantidade total de *snapshots* do grupo é dividida em dois lotes. A quantidade de *snapshots* do primeiro lote é igual ao percentual da amostragem (70), pois é maior que a quantidade mínima (30), restando um lote residual com a mesma quantidade exigida como mínima, sendo o formador do segundo lote.
- No terceiro teste a quantidade total de *snapshots* do grupo é dividida, também, em dois lotes. O primeiro lote possui a quantidade mínima exigida para processamento (40), que é a mesma definida pelo percentual (40). O segundo lote, entretanto, possui valor maior que o definido pelo percentual e maior que a quantidade mínima. Isso acontece, pois o cálculo de 40% dos 60 *snapshots* resulta em 24 *snapshots*, menor que a quantidade mínima. Assim, esse lote deveria ser formado pela quantidade mínima, ou seja, 40 *snapshots*, restando apenas 20 para o lote residual. Nesse caso, como a quantidade restante é menor que a mínima definida, todos os demais *snapshots* fazem parte do mesmo lote.

As etapas executadas pelo algoritmo para a obtenção dos resultados dos testes de mesa da Figura 62 são detalhados na Figura 63, com os indicadores sendo formados pela sequência de números obtidas por meio da concatenação dos valores de *Teste* e *Execução*, da Figura 61, formando *Teste.Execução*.

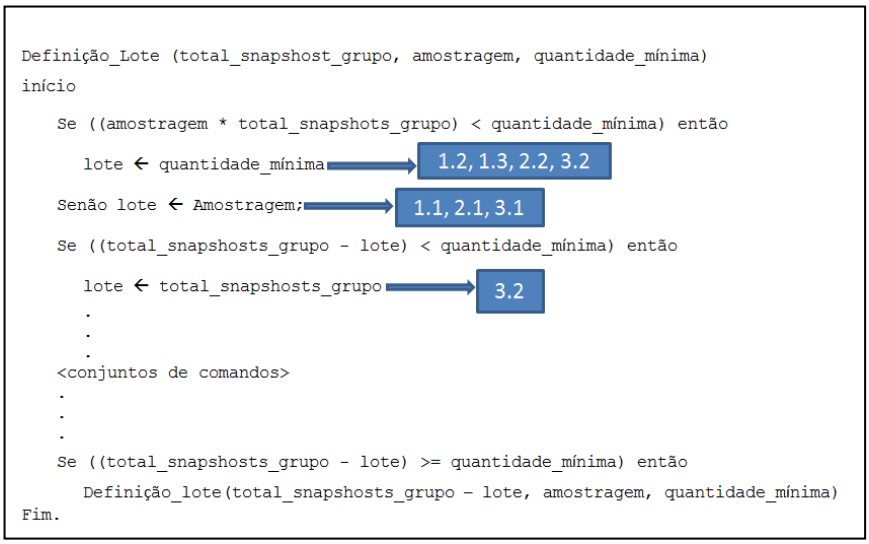

Figura 63 – Algoritmo para definição dos lotes de execução com identificação da realização dos testes de mesa

6.3 Resultados e Prioridades

Após a separação dos *snapshots* de cada grupo em lotes, inicia-se a execução individual, como instância do processo, de cada *snapshot* em um programa de um *workflow* científico. A Figura 64 esquematiza a execução desses *snapshots*.

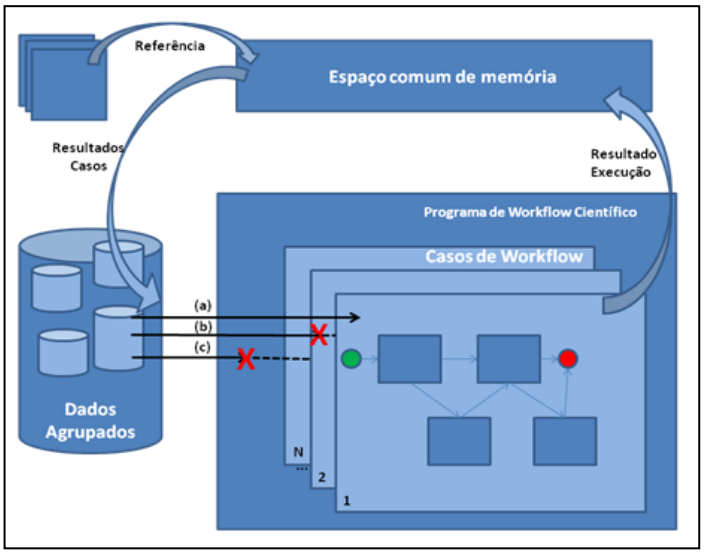

Figura 64 – Execução de *snapshots* de um grupo

Na Figura 64, (a), (b) e (c) são *snapshots* de um mesmo lote. O *snapshot* (a) é submetido à execução no programa de *workflow* científico. Como resultado dessa execução obtém-se um valor de saída, chamado de "*Resultado Execução*". Esse valor, numérico, é armazenado em um espaço comum de memória, assim chamado por ser externo ao programa de *workflow*. Esse espaço comum de memória pode ser um arquivo, armazenado em estrutura de diretórios do Sistema Operacional, ou uma tabela em um Banco de Dados. Para a análise desse valor, duas outras informações são necessárias e devem ser definidas pelo usuário, ou geradas pelo sistema por meio da execução de *snapshots* por amostragem, também representados na Figura 64 pelo identificador "*Referência*", são elas: Melhor valor e Pior valor.

Essas informações formam um intervalo [melhor_valor, pior_valor] que é utilizado para a definição de prioridades de cada um dos *snapshots*. É importante destacar que, para a aplicação estudada, considera-se como resultado do processamento de cada um dos *snapshots* o FEB, cujo resultado quanto mais negativo melhor. Devido a isso, o melhor valor é o início do intervalo e o pior valor o final do intervalo. Como exemplo, pode-se considerar o intervalo [-24, 5], considerando -24 como melhor valor e 5 como pior valor. Os resultados que se aproximarem ou forem menores que o melhor valor serão, portanto, os *snapshots* com maior probabilidade de sucesso. Busca-se processar a maior quantidade possível de *snapshots*, cujos resultados sejam

próximos do melhor valor fornecido pelo usuário. Na Figura 60 encontram-se informações referentes a resultados após a análise dos *snapshots*:

- R1', R1'', R1''', R2', R2'', R3', R4', R4'' que correspondem ao resultado final individual do processamento de cada *snapshot*;
- R11, R21, R31, R41 que correspondem ao resultado médio, obtido por meio de uma função específica, de cada um dos primeiros lotes de cada grupo.

As letras A e B da Figura 60 correspondem à possibilidade da análise dos resultados ser realizada nos dois sentidos:

- *Horizontalmente* (B), analisando o resultado de cada um dos *snapshots* de cada lote e a possibilidade de continuidade de processamento dos demais lotes de um mesmo grupo, comparando-o com o resultado médio obtido pelos demais *snapshots* do lote ao qual ele pertence;
- *Verticalmente* (A), comparando o resultado médio do processamento de um lote de um grupo com os demais lotes de outros grupos.

O algoritmo utilizado para análise dos resultados é apresentado na Figura 65.

```
calculo_resultado (numero, total_resultado, resultado_snapshot, total_lote, melhor_valor, pior_valor)
início
     numero ++;
     total_resultado ← total_resultado + resultado_snapshot;
     se (media(total_resultado,numero) e media(total_resultado,numero,(total_lote – numero),grupo,lote)) <
     media(melhor_valor,pior_valor)) então
            aumenta_prioridade_grupo();/* probabilidade de bom resultado, pois considera-se bons resultados
                                          valores quanto mais negativo melhor;*/
     senão se (media(total_resultado,numero) e media(total_resultado,numero,(total_lote –numero), grupo, lote)) >
     media(melhor_valor,pior_valor)) então
            descarta_grupo();// probabilidade de resultado ruim, pois valores estão piores que a média definida;
     senão diminui_prioridade_grupo(); // o valor está ruim, mas ainda não pode-se descartar os snapshots
fim.
```

Figura 65 – Algoritmo cálculo médio do resultado e definição prioridade

Para que seja possível analisar os resultados obtidos, utiliza-se, conforme apresentado na Figura 65, o cálculo do valor médio, obtido com o processamento dos *snapshots* do grupo em um lote, e devem ser considerados os seguintes parâmetros:

- *numero*: é a quantidade de *snapshots* já processados do lote;

- *total_resultado*: é o valor total, obtido com o somatório de todos os resultados individuais dos *snapshots* do grupo;
- *resultado_snapshot*: é o valor final de processamento do *snapshot* em questão;
- *total_lote*: é a quantidade total de *snapshots* do lote;
- *melhor_valor*: valor que serve como indicativo para o processamento, considerando esse como o melhor valor a ser atingido;
- *pior_valor*: valor que serve como indicativo para o processamento, considerando esse como o pior valor a ser atingido;

No algoritmo da Figura 65 visualiza-se, ainda, a existência da função media() com três assinaturas diferentes:

- *media (total_resultado, numero)*: retorna o valor médio, considerando-se o somatório de todos os resultados obtidos por *snapshots* já processados do lote e a quantidade de *snapshots* já processados. Por exemplo, considera-se os seguintes valores:
    - total_resultado = -100
    - numero = 25

  Calcula-se a média amostral [DEV06], considerando a Equação apresentada em (1).

$$\sum xi = \frac{\sum_{i=1}^{n} xi}{n} \quad (1)$$

  Ao se aplicar a função de média amostral, apresentada na Equação (1), com os valores já definidos tem-se a Equação (2):

$$\sum xi = \frac{-100}{25} \quad (2)$$

  Portanto a Equação (3):

$$\sum xi = -4 \quad (3)$$

- *media(melhor_valor, pior_valor)*: retorna o ponto médio entre os dois valores fornecidos, ou obtidos por meio de processamento amostral, servindo como base para a análise do resultado do lote. Consideram-se os seguintes valores exemplo para o cálculo da média aritmética, definida na Equação (1), cujos resultados são apresentados nas Equações (4) e (5):

- melhor valor: -10,8
- pior valor: -9,2

$$\sum xi = \frac{-10{,}8 + -9{,}2}{2} \quad (4)$$

$$\sum xi = -10 \quad (5)$$

- *media(total_resultado, numero, (total_lote – numero), grupo, lote)*: retorna a média amostral estimada, considerando os *snapshots* que ainda não foram processados. A expressão (total_lote – numero) fornece a quantidade de *snapshots* que ainda restam para serem processados e as informações referentes a grupo e lote são utilizadas para que se calcule o desvio padrão, resgatando os valores individuais de cada um dos *snapshots* já processados. Se for considerado o valor exemplo obtido com o cálculo da média na Equação (3), -4 está muito distante do melhor valor gerado na Equação (5), -10. Além dos valores exemplo já definidos, define-se também que total_lote é igual a 35. Portanto, apesar disso, como ainda se tem 10 *snapshots* a serem processados, considerando que 25 já foram executados, não se pode, simplesmente, inferir que o grupo não atingirá um resultado razoável se continuar seu processamento. O que se faz é verificar se o grupo tem a probabilidade de atingir e, se possível, obter um resultado médio melhor que o valor médio obtido entre o melhor e o pior valores.

  Além da média aritmética simples, é interessante que se utilize o desvio padrão para análise dos dados. Assim, o desvio padrão é utilizado para a definição da média amostral estimada, assim chamada nesse trabalho. Conforme Larson e Farber em [LAR09], a Regra Empírica, utilizada para análise dos dados por meio da utilização do desvio padrão, apresenta as características a seguir:

  - aproximadamente 68% dos dados está dentro de um desvio padrão em relação à média;
  - aproximadamente 95% dos dados está dentro de dois desvios padrão em relação à média;
  - aproximadamente 99,7% dos dados está dentro de três desvios padrão em relação à média.

Dessa forma, a Figura 66 esquematiza as características já definidas pela regra empírica, com base em [LAR09].

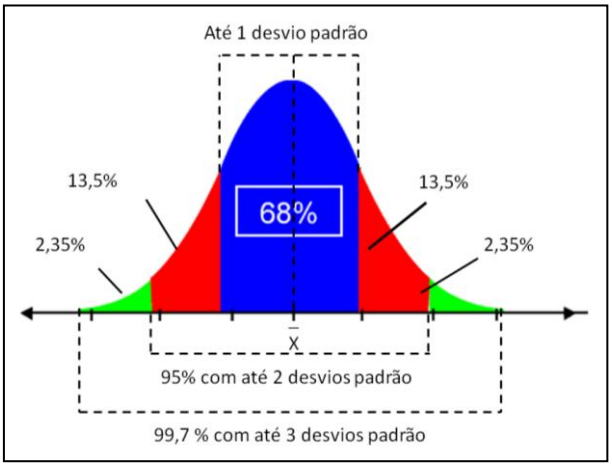

Figura 66 – Regra empírica, com base em [LAR09]

Considerando, então, a regra empírica e a análise dos desvios padrão para estimar a média dos *snapshots* faltantes, chega-se à aplicação da própria regra, obtendo o valor médio, conforme Equação (6):

$$\sum xi = \frac{\begin{array}{c}\sum_{i=1}^{n} xi + \\ \frac{\left(\left(68\%f * (m-s) + \left(68\%f * (m+s)\right)\right)\right)}{2} \\ + \\ \frac{\left(((95\%f - 68\%f) * (m-(s*2))) + ((95\%f - 68\%f) * (m+(s*2)))\right)}{2} \\ + \\ \frac{\left(((99{,}7\%f - 95\%f) * \left(m-(s*3)\right)\right) + ((99{,}7\%f - 95\%f) * \left(m+(s*3)\right))\right)}{2}\end{array}}{t} \quad (6)$$

Os valores utilizados na Equação (6) compreendem:

- f = quantidade de *snapshots* que não foram processados ainda;
- m = média já obtida com os *snapshots* processados;
- s = desvio padrão amostral, pelo processamento da população do grupo não estar concluída;
- t = quantidade total de *snapshots*, independente da quantidade já processada;
- n = quantidade de *snapshots* já processados;

Ao se aplicar a Equação (6), obtém-se um valor muito próximo da própria média já gerada, uma vez que são considerados os valores de desvio padrão e

média aritmética da própria amostra. A Tabela 4 exemplifica, com dados reais, a aplicação da Equação (6).

Tabela 4– Exemplos da aplicação da Equação definida em (6)

Média	Desvio Padrão	*Snapshots* Processados	*Snapshots* Faltantes	Média Estimada
-9,97	0,19	10	40	-9,95
-10,00	0,24	15	35	-9,98
-10,03	0,23	25	25	-10,01
-10,16	0,37	30	20	-10,15
-10,22	0,38	35	15	-10,21
-10,27	0,39	40	10	-10,26
-10,32	0,40	45	5	-10,32

Ao se selecionar a primeira linha da Tabela 4 para melhor detalhar a aplicação da Equação (6) se tem a geração dos resultados apresentados na Equação (7):

$$= \frac{((68\%40 * (-9{,}97 - 0{,}19)) + (68\%40 * (-9{,}97 + 0{,}19)))}{2}$$

$$= -271{,}3$$

$$= \frac{((((95\%40) - (68\%40)) * (-9{,}97 - (0{,}19 * 2))) + ((95\%40) - (68\%40)) * (-9{,}97 + (0{,}19 * 2)))}{2}$$

$$= -107{,}70$$

$$= \frac{(((99{,}7\%40) - (95\%40)) * (-9{,}97 - (0{,}19 * 3))) + ((99{,}7\%40) - (95\%40)) * (-9{,}97 + (0{,}19 * 3)))}{2}$$

$$= -18{,}75$$

Portanto:

$$\sum xi = \frac{\sum_{i=1}^{n} xi + (-271{,}3) + (-107{,}70) + (-18{,}75)}{t}$$

$$\sum xi = \frac{(-99{,}73) + (-271{,}3) + (-107{,}70) + (-18{,}75)}{50}$$

$$= -9{,}95$$

(7)

A diferença entre a média aritmética (-9,97) e a média estimada (-9,95) é mínima. Isso se deve ao fato de se estar utilizando a regra empírica, que apresenta a probabilidade de distribuição exata entre os valores formadores da média. Para a realidade em questão, quando se deseja analisar a probabilidade de se encontrar *snapshots* que obtenham um bom resultado, a

regra empírica não deve ser aplicada na sua forma original. Uma adaptação foi feita para que sejam superestimados os melhores valores. Assim, a regra empírica, aplicada neste trabalho para a geração da média estimada, apresenta a estrutura conforme definida pela Equação (8).

$$\sum xi = \frac{\begin{array}{c}\sum_{i=1}^{n} xi + \\ \frac{((68\%f * (m-(s*2)) + (68\%f * (m+s))}{2} \\ + \\ \frac{(((95\%f - 68\%f) * (m-(s*3))) + ((95\%f - 68\%f) * (m+(s*2))))}{2} \\ + \\ \frac{\left(((99{,}7\%f - 95\%f) * (m-(s*4))\right) + ((99{,}7\%f - 95\%f) * (m+(s*3)))))}{2}\end{array}}{t} \quad (8)$$

A principal alteração da regra utilizada neste trabalho, para a regra empírica original, está em superestimar a probabilidade de bons resultados. Assim:

- aproximadamente 68% dos *snapshots* com resultados piores (superiores nesse caso) à média estão dentro de um desvio padrão em relação à média, mas 68% dos *snapshots*, cujos resultados sejam melhores que a média (menores nesse caso) estão dentro de dois desvios em relação à média;
- aproximadamente 95% dos *snapshots* com resultados piores (superiores nesse caso) à média estão dentro de dois desvios padrão em relação à média, mas 95% dos *snapshots*, cujos resultados sejam melhores que a média (menores nesse caso) estão dentro de três desvios em relação à média;
- aproximadamente 99,7% dos *snapshots* com resultados piores (superiores nesse caso) à média estão dentro de três desvios padrão em relação à média, mas 99,7% dos *snapshots*, cujos resultados sejam melhores que a média (menores nesse caso) estão dentro de quatro desvios em relação à média.

A Figura 67 esquematiza as características da regra empírica adaptada, a qual é utilizada para o cálculo da média estimada, com base na regra empírica original [LAR09].

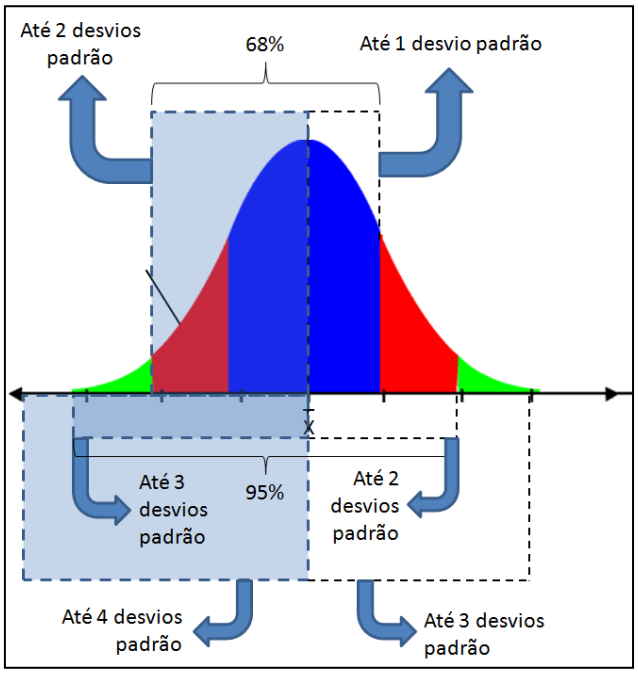

Figura 67 – Regra empírica adaptada

A Tabela 5 exemplifica a aplicação da Equação (8), que utiliza a regra empírica adaptada para a obtenção da média estimada.

Tabela 5 – Exemplos de resultados após aplicação da Equação definida em (8)

Média	**Desvio Padrão**	***Snapshots* Processados**	***Snapshots* Faltantes**	**Média Estimada**
-9,97	0,19	10	40	-10,02
-10,00	0,24	15	35	-10,07
-10,03	0,23	25	25	-10,07
-10,16	0,37	30	20	-10,22
-10,22	0,38	35	15	-10,26
-10,27	0,39	40	10	-10,30
-10,32	0,40	45	5	-10,34

Ao se selecionar a primeira linha da Tabela 5 para melhor detalhar a aplicação da Equação (8) se tem a geração dos resultados apresentados na Equação (9):

$$= \frac{((68\%40 * (-9{,}97 - (0{,}19 * 2))) + (68\%40 * (-9{,}97 + 0{,}19)))}{2}$$

$$= -273{,}9$$

$$= \frac{((((95\%40) - (68\%40)) * (-9{,}97 - (0{,}19 * 3))) + ((95\%40) - (68\%40)) * (-9{,}97 + (0{,}19 * 2)))}{2}$$

$$= -108{,}74$$

$$= \frac{(((99{,}7\%40) - (95\%40)) * (-9{,}97 - (0{,}19 * 4))) + ((99{,}7\%40) - (95\%40)) * (-9{,}97 + (0{,}19 * 3)))}{2}$$

$$= -18{,}93$$

Portanto:

$$\sum xi = \frac{\sum_{i=1}^{n} xi + (-273{,}9) + (-108{,}74) + (-18{,}93)}{t}$$

$$\sum xi = \frac{(-99{,}73) + (-273{,}9) + (-108{,}74) + (-18{,}93)}{50}$$

$$= -10{,}02$$

(9)

A média estimada é utilizada em conjunto com a média aritmética da amostra para definir a mudança de status dos *snapshots* do grupo. Portanto, busca-se que o valor gerado seja melhor em relação à média definida pelo intervalo [melhor_valor, pior_valor], pois seu resultado será utilizado para a definição do descarte ou não de *snapshots*.

Na Figura 65 as funções: *aumenta_prioridade_grupo()* e *diminui_prioridade_grupo()* são responsáveis, respectivamente, por aumentar e diminuir a prioridade do grupo ao qual o lote pertence, com base no resultado médio obtido pelo processamento dos *snapshots*. A função *descarta_grupo()* é responsável por descartar os *snapshots* do grupo, que ainda não foram processados.

O valor obtido em (9), -10,02, é melhor que o valor médio obtido na Equação (5). Portanto, considera-se, para esse caso, que o grupo tem chances de atingir um valor aceitável. Entretanto, a média amostral do grupo (-9,97) é pior que o valor médio obtido na Equação (5). Assim os *snapshots* restantes desse grupo terão sua prioridade diminuída. Essa análise deve ser feita para todos os grupos, verificando a possibilidade de que o processamento não seja realizado com todos os *snapshots*, o que reduz, consideravelmente, o tempo de execução total. É importante destacar que algumas condições devem ser satisfeitas para a aplicação do algoritmo:

1. o cálculo da média estimada será realizado quando já tiverem sido processados 20% ou mais *snapshots* do grupo;
2. para o cálculo da média estimada, utiliza-se a regra empírica adaptada, definida na Equação (8);

3. a prioridade dos *snapshots* de um grupo será mantida se sua média aritmética e sua média estimada forem melhores que a média entre o melhor e pior valores;
4. a prioridade dos *snapshots* de um grupo será reduzida se sua média aritmética for pior que a média entre o melhor e pior valores.
5. os demais *snapshots* de um grupo não serão processados se a média aritmética e a média estimada, geradas até o momento da análise, forem piores que a média entre o melhor e pior valores.

Quando o processamento estiver sendo realizado em paralelo, ou seja, mais de um grupo em processamento ao mesmo tempo, a prioridade de processamento dos lotes e *snapshots* de um grupo também pode ser alterada. Para que isso seja possível, a cada média calculada deve ser armazenada a informação em uma estrutura de arquivos/tabelas com a estrutura definida conforme Tabela 6. A Tabela 6 também apresenta alguns valores que serão utilizados para exemplificar a alteração de status de um grupo.

Tabela 6 – Estrutura de arquivo ou tabela com resultados do processamento dos lotes

Cluster	**Média**	**Média Estimada**	**Status**
0	-10,22	-10,31	A
1	-9,70	-9,86	D
2	-9,96	-10,03	P

Na Tabela 6, o valor correspondente à *Média* é o valor médio obtido entre todos os valores médios dos lotes já processados. Assim, a cada nova média aritmética obtida com o processamento de um lote, esse valor deve ser atualizado. O valor correspondente à *Média Estimada* segue o mesmo princípio: a cada nova média estimada obtida com o processamento de um lote, esse valor deve ser atualizado. O *Status* refere-se ao status do grupo como um todo. Um grupo com status igual a *P* é aquele grupo que possui a menor prioridade para processamento, ou seja, o grupo que apresenta os piores valores em *média*. Um grupo com *Status* igual a D é aquele grupo, cujos valores médios para *média* e *média estimada* são piores que o valor utilizado como parâmetro. Esse grupo foi descartado e seus *snapshots* não serão processados. Um grupo com *Status* igual a A é um grupo que está aguardando o processamento e será processado, considerando os valores obtidos pelos demais, seguindo uma ordem de prioridade, que vai do melhor valor ao pior valor. Cabe ressaltar que os status desses grupos estão sujeitos a alterações, uma vez que esses valores são atualizados e analisados a cada nova média gerada.

A diferença quando se considera status dos *snapshots* e status do grupo está na continuidade ou não de processamento de outros lotes. Se for considerado status de *snapshots* dentro de um lote após a análise da média e da média estimada para o lote, define-se:

- que os demais *snapshots* do lote serão descartados: isso significa que se passa para o processamento de outro lote e que os demais *snapshots* não serão processados.
- que os demais *snapshots* do lote terão sua prioridade reduzida: isso significa que se passa para o processamento de outro lote, mas antes de finalizar o grupo os *snapshots* que foram deixados para trás serão processados.

Se for considerado status de grupo após a análise da média e da média estimada, define-se:

- que os demais lotes do grupo serão descartados: isso significa que se passa para o processamento de outro grupo e que todos os *snapshots* dos lotes restantes não serão processados.
- que os demais lotes do grupo terão sua prioridade reduzida: isso significa que se passa para o processamento de lotes de outro grupo, mas antes de finalizar todo o processamento e reconsiderando a análise das médias, os lotes de *snapshots* que foram deixados para trás serão processados.

6.4 Considerações do Capítulo

O Padrão Múltiplas Instâncias Autoadaptáveis (P-MIA), cujo funcionamento foi apresentado neste capítulo, busca substituir as etapas envolvidas no processamento denominado como *Dados Seletivo* no *workflow* desenvolvido por Karina Machado [MAC07a], de forma que não exista a necessidade de execuções exaustivas, tornando a seleção de *snapshots* uma etapa dinâmica. Dessa forma, com o P-MIA, é possível que se analise os resultados obtidos pelos *snapshots* em tempo de execução e que se defina se *snapshots* do mesmo lote, ou do mesmo grupo, continuarão sendo executados e com qual prioridade, fornecendo ao padrão características de autoadaptação de instâncias, sem a interferência do usuário durante a execução. Estudos sobre adaptação e autoadaptação que serviram de referência para esta Tese podem ser encontrados em [HUB07] e em [HUB09].

O capítulo apresentou os algoritmos utilizados para a separação dos grupos em lotes menores, bem como as funções utilizadas para o cálculo da média aritmética e da média estimada,

cujos resultados são utilizados pelo padrão para a definição das próximas etapas de execução. Quanto à função utilizada para a média estimada, foi adaptada a Regra Empírica, definida sobre valores do desvio padrão, valor fundamental para a análise dos resultados.

No Capítulo 7, testes são apresentados, validando os algoritmos e as funções definidas.

7 P-MIA: TESTES EXPERIMENTAIS

Para a validação do funcionamento do P-MIA, apresentado no Capítulo 6, testes foram realizados com grupos, aqui também chamados de *clusters*, criados por Karina Machado, fazendo parte de pesquisas para sua Tese de Doutorado, a qual compreende a definição de uma função de similaridade específica para a realidade de estudo do LABIO. A ferramenta utilizada por Karina foi o PTRAJ que, conforme [MAC07a]:

> "é um programa utilizado para processar e analisar conjuntos de coordenadas 3D lidas de uma série de arquivos de coordenadas de entrada. Para cada conjunto de coordenadas lido, uma sequência de ações pode ser executada (em uma ordem que deve ser especificada) de acordo com configurações pré-estabelecidas".

São utilizados, nos testes desta Tese, grupos formados por dois dos algoritmos utilizados por Karina por meio do PTRAJ: *hierarchical e means (k-means)* [JAI99]. A aplicação de cada um dos algoritmos gerou 6 *clusters*, considerando os mesmos dados de entrada. Como para este trabalho a função de similaridade não é o foco parte-se, portanto, dos *clusters* já definidos para a aplicação do padrão. Os testes aqui apresentados foram realizados com dois ligantes: PIF [OLI04] e NADH [DES95].

Os testes experimentais foram realizados seguindo o processo de experimentação sugerido por Höst et al. [HOS00] e apresentado na Figura 68, [JUN10]. As próximas seções apresentam o protocolo de experimento utilizado neste trabalho, bem como o detalhamento dos testes realizados.

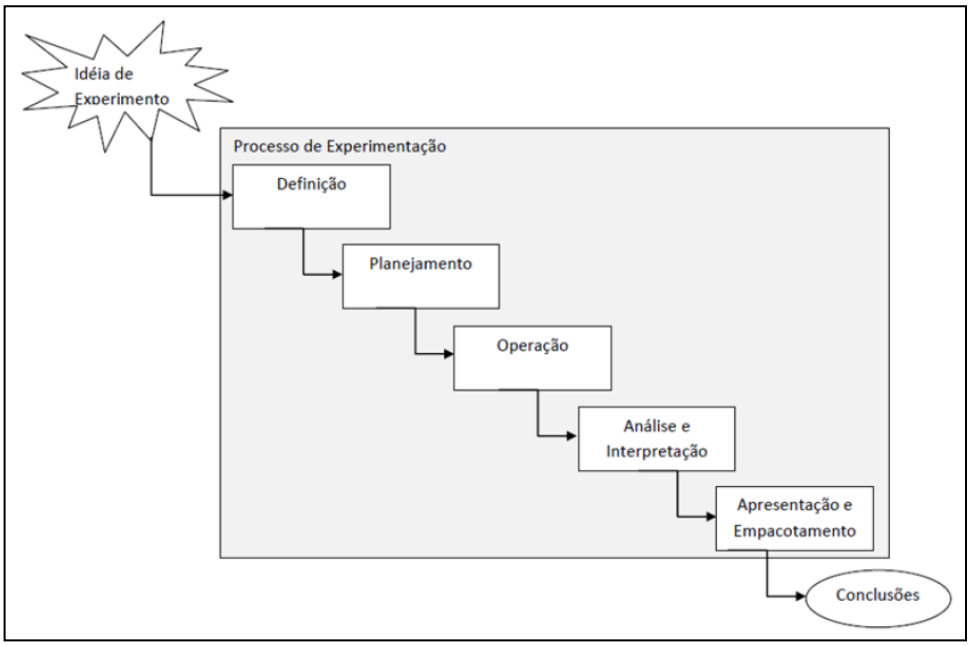

Figura 68 – Protocolo do Experimento

7.1 Definição

A *motivação* do experimento é avaliar os resultados obtidos com o P-MIA a partir dos valores gerados após a execução com *snapshots* reais. Para tanto se define que:

- O *objeto de estudo* é o resultado obtido por *snapshots* individuais após serem submetidos ao processamento em um sistema de *workflow*.
- O *propósito* do experimento é avaliar se o P-MIA pode ser utilizado pela área de Bioinformática, buscando a redução da quantidade total de *snapshots* a serem processados e garantindo que os *snapshots* que apresentariam os melhores resultados de processamento, continuariam sendo executados.
- O *foco* principal dos experimentos realizados é o desempenho do P-MIA quanto ao ganho obtido após a finalização de todas as execuções. Considera-se ganho como a quantidade de *snapshots* que não foram processados, identificados dinamicamente, sem a interferência do usuário, a possibilidade desses experimentos serem realizados em paralelo e a manutenção de resultados promissores.
- O experimento é executado no contexto de Bioinformática, mais especificamente relacionado à docagem molecular, trabalhando com dois ligantes: PIF e NADH, quando executados sobre a enzima InhA.

Portanto, os testes experimentais são realizados com o objetivo de se submeter dados reais ao P-MIA, identificando o ganho obtido.

7.2 Planejamento

O experimento é focado em analisar o funcionamento e o desempenho do P-MIA quanto ao ganho obtido após o término das execuções, por isso é classificado como *in silico*. A utilização de dados reais, no contexto experimental, conforme Junior em [JUN10], permite que outros pesquisadores possam reproduzir o experimento em questão, avaliar e sugerir melhorias ao padrão.

7.2.1 Formulação das Hipóteses

Para que os experimentos sejam avaliados coerentemente, deve-se saber como e o quê se quer formalmente avaliar. Assim, é necessário que se formule hipóteses, buscando analisar e validar o padrão proposto:

- Hipótese Nula: A utilização do P-MIA não resulta em ganhos.
- Hipótese Alternativa: A utilização do P-MIA resulta em ganhos.

Considera-se ganho como a execução de uma quantidade menor de experimentos, a possibilidade desses experimentos serem realizados em paralelo e a manutenção de resultados promissores.

7.2.2 Seleção das Variáveis

Dentre as variáveis utilizadas está o resultado do processamento de cada *snapshot*, o FEB. O FEB é utilizado, em conjunto com resultados de outros *snapshots* de um mesmo lote e de um mesmo grupo, para que se obtenha as variáveis média e média estimada, as quais subsidiam a definição de continuidade ou não do processamento. Além dessas, também são utilizados valores para amostragem e quantidade mínima de *snapshots* para serem processados, bem como, valores correspondentes ao melhor e pior valores.

7.2.3 Amostra

Os *snapshots* utilizados para execução dos experimentos apresentados nesta Tese são dos ligantes PIF e NADH. *Snapshots* de outros ligantes poderiam ter sido escolhidos, pois a amostra, em si, não define o funcionamento do padrão, o qual pode ser aplicado a qualquer conjunto de dados. Da mesma forma, foram utilizados grupos formados pelos algoritmos Hierarchical e K-Means, gerados por meio do programa PTRAJ. A diferença de utilização de

algoritmos seria representativa se fosse utilizada uma função de similaridade específica para a realidade de estudo, como a sendo proposto por Karina Machado em sua Tese de Doutorado.

7.2.4 Esboço do Experimento

Foram definidos valores para amostragem e quantidade mínima de *snapshots* para a geração dos lotes a partir dos grupos, previamente criados por uma função de similaridade qualquer. Após, com os resultados de FEB obtidos com o processamento de cada um dos *snapshots* que compõem os grupos, foram geradas médias e médias estimadas, subsidiando as análises realizadas.

7.2.5 Instrumentação

Os dados foram obtidos por meio de processamentos exaustivos realizados no LABIO. Os *snapshots* separados em grupos foram obtidos por meio de diferentes testes realizados por Karina Machado para sua Tese de Doutorado.

7.3 Operação

A preparação dos dados, para os experimentos, foi realizada associando-se os valores de FEB obtidos a seus respectivos *snapshots*, dentro de diferentes lotes, de diferentes grupos.

7.3.1 Execução

Após a separação dos *snapshots* em lotes e a associação dos resultados de FEB obtidos, os experimentos com o P-MIA validaram as funções definidas: média e média estimada para a determinação de continuidade ou não do processamento.

Os testes com os diferentes algoritmos buscam subsidiar a definição das regras já apresentadas e validar as seguintes questões:

- Grupos menores produzem resultados melhores que grupos maiores?
- Quanto antes for iniciada a análise, maior é o ganho de processamento?
- A função de similaridade aplicada está diretamente relacionada ao ganho obtido após a execução do padrão?

As próximas seções apresentam o detalhamento dos testes realizados. A etapa de análise e interpretação, da Figura 68, é apresentada em cada teste realizado. A etapa de apresentação e empacotamento, da Figura 68, não é realizada, pela característica da aplicação e as conclusões obtidas estão no final deste capítulo.

7.3.2 Testes Experimentais com o Ligante PIF

Antes da realização dos testes, algumas informações foram definidas. São elas:

- Informações necessárias para a geração dos lotes para processamento que são criados a partir dos *clusters*: quantidade mínima de *snapshots* a serem processados e amostragem (valor em percentual);
- Informações necessárias para a análise dos resultados: melhor valor e pior valor (intervalo que será considerado para a análise dos resultados).

Define-se, portanto, para a realização dos testes com o ligante PIF os seguintes valores:

- Quantidade mínima de *snapshots* para processamento por lote: 50
- Amostragem: 30%
- Melhor valor: -10,8
- Pior valor: -9,2

O intervalo gerado pelos valores fornecidos para [melhor valor, pior valor] é [-10,8; -9,2], com valor médio igual a -10. Esses valores foram obtidos por amostragem, executando-se 2 *snapshots* de cada grupo, de forma aleatória. O resultado desse processamento é apresentado na Tabela 7.

Tabela 7 – Resultado do Processamento de dois (2) *snapshots* de cada *cluster* para obter o melhor e pior valores

snapshot	***Cluster***	**FEB**
13	0	-10,08
93	0	-9,46
2719	1	-9,23
2990	1	-10,8
430	2	-9,53
460	2	-9,7
208	3	-10,03
219	3	-10,31
945	4	-10,27
1099	4	-9,95
1807	5	-9,34
1906	5	-9,2

7.3.2.1 Separação em Lotes

A Tabela 8 apresenta a quantidade de *snapshots* que compõem cada um dos grupos, gerados a partir do algoritmo *k-Means*. A Tabela 9 apresenta a definição dos lotes de

processamento de um dos *clusters* (*cluster* 0), seguindo o algoritmo da Figura 63. Os demais lotes são encontrados no Apêndice A. Para a definição dos lotes considerou-se, conforme já definido, amostragem de 30% e quantidade mínima para processamento de 50 *snapshots*.

Tabela 8 – Quantidade de *snapshots* em cada *cluster*, gerados pelo algoritmo means

Cluster	**Quantidade**
*Cluster*_0	144
*Cluster*_1	552
*Cluster*_2	484
*Cluster*_3	140
*Cluster*_4	529
*Cluster*_5	1193

Nota-se, na Tabela 9, que o lote 5 ficou maior que os lotes 2, 3 e 4. Isso se deve ao fato da amostragem definida resultar em quantidade menor que a quantidade mínina de *snapshots* exigida para processamento e ao lote residual final ficar com menos *snapshots* que o mínimo definido.

Tabela 9 – Lotes para processamento do *cluster* 1

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
1	0	166	1595 1596 1597 1613 1619 1681 1692 1698 1714 1715 1716 1717 1726 1732 1768 1836 1837 1852 2018 2073 2108 2123 2152 2158 2159 2162 2164 2166 2167 2168 2223 2227 2233 2235 2237 2239 2241 2242 2244 2249 2250 2251 2254 2255 2256 2257 2258 2259 2260 2264 2265 2266 2267 2268 2300 2405 2407 2408 2410 2411 2413 2414 2429 2440 2441 2451 2455 2458 2459 2504 2515 2522 2543 2545 2546 2549 2550 2551 2552 2556 2557 2558 2559 2565 2568 2571 2572 2577 2580 2582 2583 2585 2586 2588 2589 2590 2591 2592 2593 2594 2595 2596 2598 2599 2600 2602 2603 2604 2605 2607 2608 2609 2610 2613 2614 2615 2617 2618 2619 2620 2621 2622 2623 2624 2625 2626 2627 2628 2629 2630 2631 2632 2633 2634 2635 2637 2638 2639 2640 2641 2642 2643 2644 2645 2646 2647 2648 2649 2650 2652 2653 2654 2655 2656 2657 2658 2659 2660 2661 2662 2663 2664 2666 2667 2668 2669
1	1	116	2670 2671 2672 2673 2674 2676 2677 2678 2679 2680 2681 2682 2683 2690 2691 2692 2693 2694 2695 2696 2697 2703 2704 2705 2706 2707 2708 2709 2712 2713 2714 2715 2716 2717 2718 2719 2720 2721 2722 2723 2725 2726 2727 2728 2729 2730 2731 2732 2733 2734 2735 2736 2737 2738 2739 2740 2741 2743 2744 2745 2746 2747 2748 2749 2750 2751 2753 2754 2755 2756 2757 2758 2759 2760 2761 2762 2763 2764 2765 2766 2767 2768 2769 2770 2771 2772 2774 2775 2776 2777 2778 2779 2780 2781 2782 2783 2784 2785 2786 2787 2788 2789 2790 2791 2792 2793 2794 2795 2796 2797 2798 2799 2801 2802 2803 2804
1	2	81	2805 2806 2807 2808 2809 2810 2811 2812 2813 2814 2815 2816 2817 2818 2819 2820 2821 2822 2823 2824 2825 2826 2827 2828 2829 2830 2831 2832 2833 2834 2835 2836 2837 2838 2839 2840 2841 2842 2843 2844 2845 2846 2847 2848 2849 2850 2851 2852 2853 2854 2855 2856 2857 2858 2859 2860 2861 2862 2863 2864 2865 2866 2867 2868 2869 2870 2871 2872 2873 2874 2875 2876 2877 2878 2879 2880 2881 2882 2884 2885 2886
1	3	57	2887 2888 2889 2890 2891 2892 2893 2894 2895 2896 2897 2898 2899 2900 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 2911 2912 2913 2914 2915 2916 2917 2919 2920 2921 2922 2923 2924 2925 2926 2927 2928 2929 2930 2931 2932 2933 2934 2935 2936 2937 2938 2939 2940 2941 2942 2943 2944

1	4	50	2945 2946 2947 2948 2949 2950 2951 2952 2953 2954 2955 2956 2957 2958 2959 2960 2961 2962 2963 2964 2965 2966 2967 2968 2969 2970 2971 2975 2976 2978 2979 2980 2984 2985 2986 2988 2989 2990 2993 2994 2995 2996 2998 2999 3000 3001 3002 3003 3004 3005
1	5	82	3006 3007 3008 3009 3010 3011 3012 3013 3014 3015 3016 3017 3018 3019 3020 3021 3022 3023 3024 3025 3026 3027 3028 3030 3031 3032 3033 3034 3035 3036 3037 3038 3039 3040 3041 3042 3043 3044 3045 3046 3047 3048 3049 3050 3051 3052 3053 3054 3056 3057 3058 3059 3060 3061 3062 3063 3064 3065 3066 3068 3069 3070 3071 3072 3073 3074 3075 3076 3077 3086 3088 3090 3091 3092 3093 3094 3095 3096 3097 3098 3099 3100

7.3.2.2 Resultados Obtidos

Após a separação dos *snapshots* do *cluster* 1 em lotes inicia-se a execução individual, como instância do processo, de cada um dos *snapshots*. A Tabela 10 apresenta 3 *snapshots* já processados, os resultados individuais de processamento obtidos e os próximos *snapshots* a serem processados.

Tabela 10 – *Snapshots* já processados e *snapshots* aguardando o processamento

Snapshot	***Cluster***	**Lote**	**Status**	**Resultado**
1595	1	0	F	-9,96
1596	1	0	F	-9,45
1597	1	0	F	-9,9
1613	1	0	A	
1619	1	0	A	
1681	1	0	A	

Conforme especificado no Capítulo 6, os valores possíveis para o status dos *snapshots* são: A (Ativo); F (Finalizado); D (Descartado); P (Prioridade Alterada – diminuída). Para um lote com 166 *snapshots*, o processamento de 3 *snapshots* não permite a realização de muitas análises, dessa forma, processam-se mais *snapshots* antes que seja possível se inferir algumas conclusões.

Testes foram realizados, com quantidades diferentes de *snapshots*, buscando identificar qual seria o percentual a ser utilizado para definição de continuidade ou descarte do lote. É importante destacar que a definição de alteração de prioridade e descarte deve ser analisada, conforme já mencionado, com base na média aritmética dos valores obtidos e na média estimada. As Tabelas 11, 12, 13, 14 e 15 apresentam as análises realizadas. Para o correto entendimento das informações contidas nas tabelas as colunas referem-se a: (i) lote de cada *cluster*: identificado por *C_L*; (ii) quantidade total de *snapshots* do lote: identificado por *Quant*; (iii) média aritmética dos *snapshots* do lote até o momento da análise: identificado por *M20%, M30%, M50%, M70%* e *M80%*; (iv) média estimada dos *snapshots* restantes até o momento da análise: identificado por *E20%, E30%, E50%, E70%* e *E80%*; (v) quantidade de *snapshots* processados até o momento da

análise: identificado por *Proc*; (vi) quantidade total de *snapshots* processados: identificado por *ProcFinal*; e (vii) quantidade de *snapshots* que não precisaram ser processados: identificados por *Ganho*.

Uma característica em comum nas Tabelas 11, 12, 13, 14 e 15 é a presença de sombreamentos. Os sombreamentos à direita das colunas que estão sendo analisadas significam que os *snapshots* em questão não foram processados.

Tabela 11 – Análise dos resultados com 20% dos *snapshots* processados

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	Proc Final	Ganho
C0_L0	50	-10,0	-10,0	10	-10,0	-10,1	5	-10,0	-10,07	15	-10,22	-10,3	5	-10,27	-10,30	40	50	0
C0_L1	94	-10,5	-10,7	19	-10,5	-10,55	9	-10,3	-10,39	19	-10,24	-10,3	19	-10,18	-10,20	75	94	0
C1_L0	166	-10,0	-10,1	33	-10,1	-10,19	17	-10,0	-10,06	33	-9,86	-9,9	33				116	50
C1_L1	116	-9,19	-9,64	23													23	93
C1_L2	81	-9,80	-9,86	16													16	65
C1_L3	57	-9,78	-9,82	11													11	46
C1_L4	50	-9,74	-9,90	10													10	40
C1_L5	82	-9,91	-9,95	16													16	66
C2_L0	145	-10,10	-10,20	29	-10,12	-10,23	16	-10,01	-10,07	28	-9,96	-10,03	29	-9,96	-9,99	14	116	29
C2_L1	102	-9,87	-9,95	20													20	82
C2_L2	71	-10,01	-10,05	14	-10,04	-10,07	7	-10,04	-10,10	14	-10,03	-10,10	15	-10,06	-10,11	7	71	0
C2_L3	50	-9,63	-10,22	10	-9,77	-10,06	5	-9,74	-10,02	10	-9,83	-9,97	10				35	15
C2_L4	50	-9,92	-10,01	10	-9,93	-10,01	5	-9,94	-10,00	10	-9,95	-9,98	10				35	15
C2_L5	66	-10,00	-10,07	13	-10,05	-10,17	6	-10,02	-10,09	14	-9,99	-10,01	13	-9,98	-10,03	7	66	0
C3_L0	50	-9,98	-10,09	10	-10,02	-10,09	5	-10,05	-10,11	10	-10,05	-10,09	10	-10,05	-10,07	5	50	0
C3_L1	90	-10,10	-10,17	18	-10,11	-10,17	9	-10,12	-10,17	18	-10,12	-10,15	18	-10,11	-10,13	9	90	0
C4_L0	159	-10,02	-10,19	32	-10,08	-10,22	48	-10,11	-10,23	32	-10,11	-10,13	31	-10,10	-10,12	16	159	0
C4_L1	111	-9,97	-10,05	22	-9,99	-10,03	11										33	78
C4_L2	78	-9,71	-9,83	16													16	62
C4_L3	54	-9,70	-9,84	11													11	43
C4_L4	50	-10,11	-10,27	10	-10,06	-10,19	5	-10,07	-10,14	10	-10,00	-10,05	10	-9,98	-10,01	5	50	0
C4_L5	77	-9,77	-9,84	15													15	62
C5_L0	356	-9,95	-10,00	71	-9,89	-9,95	36										107	249
C5_L1	251	-9,69	-9,80	50													50	201
C5_L2	176	-10,17	-10,29	35	-10,10	-10,23	18	-10,15	-10,24	35	-10,13	-10,18	35	-10,13	-10,18	18	176	0
C5_L3	123	-10,16	-10,32	25	-10,19	-10,30	12	-10,14	-10,27	24	-10,09	-10,13	25	-10,10	-10,10	12	123	0
C5_L4	86	-9,67	-9,78	17													17	69
C5_L5	60	-9,50	-9,57	12													12	48
C5_L6	50	-9,59	-9,65	10													10	40
C5_L7	91	-9,71	-9,82	18													18	73
Ganho de Processamento – *snapshots*																		1426
Ganho de Processamento – percentual																		47%

Na Tabela 11, a análise dos resultados da média aritmética e da média estimada começou quando 20% dos *snapshots* foram processados. Essa análise deve ser refeita para os lotes que, quando analisados em 20% de processamento, continuaram com *snapshots* com status de Ativo, ou Prioridade Reduzida. Os lotes, cujos *snapshots* passaram a ter status de Descartado não serão mais analisados. Esse status é definido quando os valores de média e média estimada, ambos, são

piores que o valor médio utilizado como parâmetro. Um exemplo pode ser visualizado na Tabela 11, no *cluster* 1, Lote 1 (C1_L1), quando a média possui valor igual a -9,19 e a média estimada igual a -9,64 (valor médio igual a -10). Aos *snapshots* que apresentam essa característica, após esse ponto de análise, foi conferido o sombreamento nas tabelas, sem resultado numérico na célula.

Tabela 12 – Análise dos resultados com 30% dos *snapshots* processados

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	Proc Final	Ganho
C0_L0	50	-10,0	-10,0	10	-10,0	-10,1	15	-10,03	-10,07	15	-10,22	-10,26	5	-10,27	-10,30	40	50	0
C0_L1	94	-10,5	-10,7	19	-10,5	-10,6	28	-10,31	-10,39	19	-10,24	-10,31	19	-10,18	-10,20	75	94	0
C1_L0	166	-10,0	-10,1	33	-10,1	-10,2	50	-9,97	-10,06	33	-9,86	-9,94	33				116	50
C1_L1	116	-9,19	-9,64	23	-9,40	-9,81	35										35	81
C1_L2	81	-9,80	-9,86	16	-9,73	-9,79	24										24	57
C1_L3	57	-9,78	-9,82	11	-9,84	-9,91	17										17	40
C1_L4	50	-9,74	-9,90	10	-9,81	-9,92	15										15	35
C1_L5	82	-9,91	-9,95	16	-9,93	-10,07	25	-10,00	-10,07	16	-10,02	-10,01	16	-10,02	-10,10	9	82	0
C2_L0	145	-10,10	-10,20	29	-10,12	-10,23	44	-10,01	-10,07	28	-9,96	-10,03	29	-9,96	-9,99	14	145	0
C2_L1	102	-9,87	-9,95	20	-9,84	-9,96	31										31	71
C2_L2	71	-10,01	-10,05	14	-10,04	-10,07	21	-10,04	-10,10	14	-10,03	-10,10	15	-10,06	-10,11	7	71	0
C2_L3	50	-9,63	-10,22	10	-9,77	-10,06	15	-9,74	-10,02	10	-9,83	-9,97	10				35	15
C2_L4	50	-9,92	-10,01	10	-9,93	-10,01	15	-9,94	-10,00	10	-9,95	-9,98	10				35	15
C2_L5	66	-10,00	-10,07	13	-10,05	-10,17	20	-10,02	-10,09	14	-9,99	-10,01	13	-9,98	-10,03	7	66	0
C3_L0	50	-9,98	-10,09	10	-10,02	-10,09	15	-10,05	-10,11	10	-10,05	-10,09	10	-10,05	-10,07	5	50	0
C3_L1	90	-10,10	-10,17	18	-10,11	-10,17	27	-10,12	-10,17	18	-10,12	-10,15	18	-10,11	-10,13	9	90	0
C4_L0	159	-10,02	-10,19	32	-10,08	-10,22	48	-10,11	-10,23	32	-10,11	-10,13	31	-10,10	-10,12	16	159	0
C4_L1	111	-9,97	-10,05	22	-9,99	-10,03	33										33	78
C4_L2	78	-9,71	-9,83	16	-9,72	-9,74	23										23	55
C4_L3	54	-9,70	-9,84	11	-9,75	-9,80	16										16	38
C4_L4	50	-10,11	-10,27	10	-10,06	-10,19	15	-10,07	-10,14	10	-10,00	-10,05	10	-9,98	-10,01	5	50	0
C4_L5	77	-9,77	-9,84	15	-9,80	-9,87	23										23	54
C5_L0	356	-9,95	-10,00	71	-9,89	-9,95	107										107	249
C5_L1	251	-9,69	-9,80	50	-9,79	-9,85	75										75	176
C5_L2	176	-10,17	-10,29	35	-10,10	-10,23	53	-10,15	-10,24	35	-10,13	-10,18	35	-10,13	-10,18	18	176	0
C5_L3	123	-10,16	-10,32	25	-10,19	-10,30	37	-10,14	-10,27	24	-10,09	-10,13	25	-10,10	-10,10	12	123	0
C5_L4	86	-9,67	-9,78	17	-9,67	-9,80	26										26	60
C5_L5	60	-9,50	-9,57	12	-9,56	-9,62	18										18	42
C5_L6	50	-9,59	-9,65	10	-9,63	-9,70	15										15	35
C5_L7	91	-9,71	-9,82	18	-9,73	-9,79	27										27	64
Ganho de Processamento - *snapshots*																		1215
Ganho de Processamento - percentual																		40%

Iniciando a análise, conforme apresentado na Tabela 11, com 20% de processamento, obtém-se um ganho para os dados analisados de 47%, ou seja, 1426 *snapshots* de um total de 3042 não foram processados, considerando-se a média aritmética e a média estimada. Uma característica interessante do modelo, que pode ser comprovada por meio da Tabela 11 nas linhas correspondentes ao *Cluster* 2, Lotes 3 e 4 (C2_L3, C2_L4) é que a média dos lotes (-9,63 e -9,92) é pior que a média que está sendo utilizada como parâmetro (-10,0). Apesar disso, os *snapshots* continuaram sendo processados, com prioridade reduzida, pois as médias estimadas dos lotes são maiores que a média utilizada como parâmetro (-10,22 e -10,01). *Snapshots* desses lotes somente

deixam de ser processados quando já tiver sido executado 70% de cada um dos lotes. Nesse ponto, as médias estimadas tornam-se piores que a média utilizada como parâmetro (-9,97 e -9,98).

Tabela 13 – Análise dos resultados com 50% dos *snapshots* processados

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	Proc Final	Ganho
C0_L0	50	-10,0	-10,0	10	-10,0	-10,1	15	-10,0	-10,1	25	-10,2	-10,3	5	-10,27	-10,30	40	50	0
C0_L1	94	-10,5	-10,7	19	-10,5	-10,6	28	-10,3	-10,4	47	-10,24	-10,3	19	-10,18	-10,20	75	94	0
C1_L0	166	-10,0	-10,1	33	-10,1	-10,2	50	-10,0	-10,1	83	-9,86	-9,9	33				116	50
C1_L1	116	-9,19	-9,64	23	-9,40	-9,81	35	-9,50	-9,78	58							58	58
C1_L2	81	-9,80	-9,86	16	-9,73	-9,79	24	-9,7	-9,7	41							41	40
C1_L3	57	-9,78	-9,82	11	-9,84	-9,91	17	-9,85	-9,92	29							29	28
C1_L4	50	-9,74	-9,90	10	-9,81	-9,92	15	-9,78	-9,89	25							25	25
C1_L5	82	-9,91	-9,95	16	-9,93	-10,07	25	-10,00	-10,07	41	-10,02	-10,01	-9	-10,02	-10,10	9	82	0
C2_L0	145	-10,10	-10,20	29	-10,12	-10,23	44	-10,01	-10,07	73	-9,96	-10,03	29	-9,96	-9,99	14	145	0
C2_L1	102	-9,87	-9,95	20	-9,84	-9,96	31	-9,81	-9,87	51							51	51
C2_L2	71	-10,01	-10,05	14	-10,04	-10,07	21	-10,04	-10,10	36	-10,03	-10,10	15	-10,06	-10,11	7	71	0
C2_L3	50	-9,63	-10,22	10	-9,77	-10,06	15	-9,74	-10,02	25	-9,83	-9,97	10				35	15
C2_L4	50	-9,92	-10,01	10	-9,93	-10,01	15	-9,94	-10,00	25	-9,95	-9,98	10				35	15
C2_L5	66	-10,00	-10,07	13	-10,05	-10,17	20	-10,02	-10,09	33	-9,99	-10,01	13	-9,98	-10,03	7	66	0
C3_L0	50	-9,98	-10,09	10	-10,02	-10,09	15	-10,05	-10,11	25	-10,05	-10,09	10	-10,05	-10,07	5	50	0
C3_L1	90	-10,10	-10,17	18	-10,11	-10,17	27	-10,12	-10,17	45	-10,12	-10,15	18	-10,11	-10,13	9	90	0
C4_L0	159	-10,02	-10,19	32	-10,08	-10,22	48	-10,11	-10,23	80	-10,11	-10,13	31	-10,10	-10,12	16	159	0
C4_L1	111	-9,97	-10,05	22	-9,99	-10,03	33	-9,87	-9,99	56							56	55
C4_L2	78	-9,71	-9,83	16	-9,72	-9,74	23	-9,74	-9,79	39							39	39
C4_L3	54	-9,70	-9,84	11	-9,75	-9,80	16	-9,75	-9,82	27							27	27
C4_L4	50	-10,11	-10,27	10	-10,06	-10,19	15	-10,07	-10,14	25	-10,00	-10,05	10	-9,98	-10,01	5	50	0
C4_L5	77	-9,77	-9,84	15	-9,80	-9,87	23	-9,84	-9,97	39							39	38
C5_L0	356	-9,95	-10,00	71	-9,89	-9,95	107	-9,86	-9,93	178							178	178
C5_L1	251	-9,69	-9,80	50	-9,79	-9,85	75	-9,68	-9,78	126							126	125
C5_L2	176	-10,17	-10,29	35	-10,10	-10,23	53	-10,15	-10,24	88	-10,13	-10,18	35	-10,13	-10,18	18	176	0
C5_L3	123	-10,16	-10,32	25	-10,19	-10,30	37	-10,14	-10,27	62	-10,09	-10,13	25	-10,10	-10,10	12	123	0
C5_L4	86	-9,67	-9,78	17	-9,67	-9,80	26	-9,71	-9,78	43							43	43
C5_L5	60	-9,50	-9,57	12	-9,56	-9,62	18	-9,55	-9,60	30							30	30
C5_L6	50	-9,59	-9,65	10	-9,63	-9,70	15	-9,63	-9,69	25							25	25
C5_L7	91	-9,71	-9,82	18	-9,73	-9,79	27	-9,67	-9,80	46							46	45
Ganho de Processamento - *snapshots*																		887
Ganho de Processamento - percentual																		29%

Na Tabela 12, as análises começaram a ser feitas quando 30% do total de *snapshots* de cada lote estavam processados. Assim, os cálculos de média aritmética e média estimada foram realizados, considerando os *snapshots* executados. Ao se confrontar a Tabela 11 com a Tabela 12, verifica-se que o Lote 5 do *Cluster* 1 apresenta diferenças consideráveis de ação. Na Tabela 11, os demais *snapshots* desse lote, após o processamento de 20%, foram descartados, pois a média aritmética (-9,91) e a média estimada (-9,95) são piores que o valor médio utilizado como parâmetro (-10,0). Já na Tabela 12, quando a análise foi iniciada após 30% do processamento do lote concluído, os demais *snapshots* são processados, com baixa prioridade, mas processados, pois o valor da média aritmética (-9,93) é pior que o valor utilizado como parâmetro, mas a média

estimada (-10,07) possui valor melhor. Os demais lotes seguiram a mesma tendência identificada na Tabela 11. Com a análise iniciando com 30% do processamento concluído, obtém-se um ganho de 40% dos *snapshots*, ou seja, 1215 *snapshots*, dos 3042, não foram processados.

Na Tabela 13 a análise dos resultados foi iniciada após ter sido concluído o processamento de 50% do total de *snapshots* do lote. Não se evidencia mudanças da estratégia de execução dos *snapshots* quando analisados em 50% (Tabela 13) ou analisados em 30% (Tabela 12). Como a análise foi postergada, o ganho de processamento, nesse caso, é menor. Quando se inicia a análise para inferir sobre os status dos *snapshots* com 50% de processamento concluído, obtém-se um ganho de 29%, ou seja, 887 *snapshots* não foram processados.

Tabela 14 – Análise dos resultados com 70% dos *snapshots* processados

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	Proc Final	Ganho
C0_L0	50	-10,0	-10,0	10	-10,0	-10,1	15	-10,0	-10,1	25	-10,22	-10,3	35	-10,27	-10,30	40	50	0
C0_L1	94	-10,5	-10,7	19	-10,5	-10,6	28	-10,3	-10,4	47	-10,24	-10,3	66	-10,18	-10,20	75	94	0
C1_L0	166	-10,0	-10,1	33	-10,1	-10,2	50	-10,0	-10,1	83	-9,86	-9,9	116				116	50
C1_L1	116	-9,19	-9,64	23	-9,40	-9,81	35	-9,50	-9,78	58	-9,53	-9,66	81				81	35
C1_L2	81	-9,80	-9,86	16	-9,73	-9,79	24	-9,7	-9,7	41	-9,64	-9,72	57				57	24
C1_L3	57	-9,78	-9,82	11	-9,84	-9,91	17	-9,85	-9,92	29	-9,77	-9,84	40				40	17
C1_L4	50	-9,74	-9,90	10	-9,81	-9,92	15	-9,78	-9,89	25	-9,89	-9,95	35				35	15
C1_L5	82	-9,91	-9,95	16	-9,93	-10,07	25	-10,00	-10,07	41	-10,02	-10,01	57	-10,02	-10,10	9	82	0
C2_L0	145	-10,10	-10,20	29	-10,12	-10,23	44	-10,01	-10,07	73	-9,96	-10,03	102	-9,96	-9,99	14	145	0
C2_L1	102	-9,87	-9,95	20	-9,84	-9,96	31	-9,81	-9,87	51	-9,77	-9,80	71				71	31
C2_L2	71	-10,01	-10,05	14	-10,04	-10,07	21	-10,04	-10,10	36	-10,03	-10,10	50	-10,06	-10,11	7	71	0
C2_L3	50	-9,63	-10,22	10	-9,77	-10,06	15	-9,74	-10,02	25	-9,83	-9,97	35				35	15
C2_L4	50	-9,92	-10,01	10	-9,93	-10,01	15	-9,94	-10,00	25	-9,95	-9,98	35				35	15
C2_L5	66	-10,00	-10,07	13	-10,05	-10,17	20	-10,02	-10,09	33	-9,99	-10,01	46	-9,98	-10,03	7	66	0
C3_L0	50	-9,98	-10,09	10	-10,02	-10,09	15	-10,05	-10,11	25	-10,05	-10,09	35	-10,05	-10,07	5	50	0
C3_L1	90	-10,10	-10,17	18	-10,11	-10,17	27	-10,12	-10,17	45	-10,12	-10,15	63	-10,11	-10,13	9	90	0
C4_L0	159	-10,02	-10,19	32	-10,08	-10,22	48	-10,11	-10,23	80	-10,11	-10,13	111	-10,10	-10,12	16	159	0
C4_L1	111	-9,97	-10,05	22	-9,99	-10,03	33	-9,87	-9,99	56	-9,87	-9,94	78				78	33
C4_L2	78	-9,71	-9,83	16	-9,72	-9,74	23	-9,74	-9,79	39	-9,75	-9,84	55				55	23
C4_L3	54	-9,70	-9,84	11	-9,75	-9,80	16	-9,75	-9,82	27	-9,85	-9,93	38				38	16
C4_L4	50	-10,11	-10,27	10	-10,06	-10,19	15	-10,07	-10,14	25	-10,00	-10,05	35	-9,98	-10,01	5	50	0
C4_L5	77	-9,77	-9,84	15	-9,80	-9,87	23	-9,84	-9,97	39	-9,86	-9,91	54				54	23
C5_L0	356	-9,95	-10,00	71	-9,89	-9,95	107	-9,86	-9,93	178	-9,84	-9,88	249				249	107
C5_L1	251	-9,69	-9,80	50	-9,79	-9,85	75	-9,68	-9,78	126	-9,66	-9,72	176				176	75
C5_L2	176	-10,17	-10,29	35	-10,10	-10,23	53	-10,15	-10,24	88	-10,13	-10,18	123	-10,13	-10,18	18	176	0
C5_L3	123	-10,16	-10,32	25	-10,19	-10,30	37	-10,14	-10,27	62	-10,09	-10,13	86	-10,10	-10,10	12	123	0
C5_L4	86	-9,67	-9,78	17	-9,67	-9,80	26	-9,71	-9,78	43	-9,72	-9,74	60				60	26
C5_L5	60	-9,50	-9,57	12	-9,56	-9,62	18	-9,55	-9,60	30	-9,54	-9,57	42				42	18
C5_L6	50	-9,59	-9,65	10	-9,63	-9,70	15	-9,63	-9,69	25	-9,67	-9,71	35				35	15
C5_L7	91	-9,71	-9,82	18	-9,73	-9,79	27	-9,67	-9,80	46	-9,49	-9,64	64				64	27
Ganho de Processamento - *snapshots*																		565
Ganho de Processamento – percentual																		19%

Na Tabela 14, da mesma forma que o analisado para a Tabela 13, não existem mudanças a serem consideradas. Há, apenas, a redução no ganho de processamento por se iniciar a análise

após 70% dos *snapshots* do lote terem sido processados. O ganho, nesse caso, foi de 19%, ou seja, 565 *snapshots* não foram processados.

Na Tabela 15, quando se inicia a análise após 80% dos *snapshots* terem sido processados, a redução é menor, apenas de 13%, ou seja, 404 *snapshots* não precisaram ser processados pela característica do lote, gerada por meio dos resultados obtidos pelos *snapshots* já processados.

Tabela 15 – Análise dos resultados com 80% dos *snapshots* processados

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	Proc Final	Ganho
C0_L0	50	-10,0	-10,0	10	-10,0	-10,1	15	-10,0	-10,1	25	-10,22	-10,3	35	-10,27	-10,30	40	50	0
C0_L1	94	-10,5	-10,7	19	-10,5	-10,6	28	-10,3	-10,4	47	-10,24	-10,3	66	-10,18	-10,20	75	94	0
C1_L0	166	-10,0	-10,1	33	-10,1	-10,2	50	-10,0	-10,1	83	-9,86	-9,9	116	-9,83	-9,92	133	133	33
C1_L1	116	-9,19	-9,64	23	-9,40	-9,81	35	-9,50	-9,78	58	-9,53	-9,66	81	-9,58	-9,68	93	93	23
C1_L2	81	-9,80	-9,86	16	-9,73	-9,79	24	-9,7	-9,7	41	-9,64	-9,72	57	-9,65	-9,70	65	65	16
C1_L3	57	-9,78	-9,82	11	-9,84	-9,91	17	-9,85	-9,92	29	-9,77	-9,84	40	-9,77	-9,87	46	46	11
C1_L4	50	-9,74	-9,90	10	-9,81	-9,92	15	-9,78	-9,89	25	-9,89	-9,95	35	-9,89	-9,93	40	40	10
C1_L5	82	-9,91	-9,95	16	-9,93	-10,07	25	-10,00	-10,07	41	-10,02	-10,01	57	-10,02	-10,10	66	82	0
C2_L0	145	-10,10	-10,20	29	-10,12	-10,23	44	-10,01	-10,07	73	-9,96	-10,03	102	-9,96	-9,99	116	116	29
C2_L1	102	-9,87	-9,95	20	-9,84	-9,96	31	-9,81	-9,87	51	-9,77	-9,80	71	-9,83	-9,91	82	82	20
C2_L2	71	-10,01	-10,05	14	-10,04	-10,07	21	-10,04	-10,10	36	-10,03	-10,10	50	-10,06	-10,11	57	71	0
C2_L3	50	-9,63	-10,22	10	-9,77	-10,06	15	-9,74	-10,02	25	-9,83	-9,97	35	-9,85	-9,94	40	40	10
C2_L4	50	-9,92	-10,01	10	-9,93	-10,01	15	-9,94	-10,00	25	-9,95	-9,98	35	-9,97	-10,00	40	40	10
C2_L5	66	-10,00	-10,07	13	-10,05	-10,17	20	-10,02	-10,09	33	-9,99	-10,01	46	-9,98	-10,03	53	66	0
C3_L0	50	-9,98	-10,09	10	-10,02	-10,09	15	-10,05	-10,11	25	-10,05	-10,09	35	-10,05	-10,07	40	50	0
C3_L1	90	-10,10	-10,17	18	-10,11	-10,17	27	-10,12	-10,17	45	-10,12	-10,15	63	-10,11	-10,13	72	90	0
C4_L0	159	-10,02	-10,19	32	-10,08	-10,22	48	-10,11	-10,23	80	-10,11	-10,13	111	-10,10	-10,12	127	159	0
C4_L1	111	-9,97	-10,05	22	-9,99	-10,03	33	-9,87	-9,99	56	-9,87	-9,94	78	-9,87	-9,92	89	89	22
C4_L2	78	-9,71	-9,83	16	-9,72	-9,74	23	-9,74	-9,79	39	-9,75	-9,84	55	-9,75	-9,73	62	62	16
C4_L3	54	-9,70	-9,84	11	-9,75	-9,80	16	-9,75	-9,82	27	-9,85	-9,93	38	-9,85	-9,84	43	43	11
C4_L4	50	-10,11	-10,27	10	-10,06	-10,19	15	-10,07	-10,14	25	-10,00	-10,05	35	-9,98	-10,01	40	50	0
C4_L5	77	-9,77	-9,84	15	-9,80	-9,87	23	-9,84	-9,97	39	-9,86	-9,91	54	-9,82	-9,90	62	62	15
C5_L0	356	-9,95	-10,00	71	-9,89	-9,95	107	-9,86	-9,93	178	-9,84	-9,88	249	-9,83	-9,86	285	285	71
C5_L1	251	-9,69	-9,80	50	-9,79	-9,85	75	-9,68	-9,78	126	-9,66	-9,72	176	-9,63	-9,67	201	201	50
C5_L2	176	-10,17	-10,29	35	-10,10	-10,23	53	-10,15	-10,24	88	-10,13	-10,18	123	-10,13	-10,18	141	176	0
C5_L3	123	-10,16	-10,32	25	-10,19	-10,30	37	-10,14	-10,27	62	-10,09	-10,13	86	-10,10	-10,10	98	123	0
C5_L4	86	-9,67	-9,78	17	-9,67	-9,80	26	-9,71	-9,78	43	-9,72	-9,74	60	-9,73	-9,78	69	69	17
C5_L5	60	-9,50	-9,57	12	-9,56	-9,62	18	-9,55	-9,60	30	-9,54	-9,57	42	-9,57	-9,59	48	48	12
C5_L6	50	-9,59	-9,65	10	-9,63	-9,70	15	-9,63	-9,69	25	-9,67	-9,71	35	-9,69	-9,72	40	40	10
C5_L7	91	-9,71	-9,82	18	-9,73	-9,79	27	-9,67	-9,80	46	-9,49	-9,64	64	-9,56	-9,66	73	73	18
Ganho de Processamento - *snapshots*																		404
Ganho de Processamento - percentual																		13%

7.3.2.3 Análise e Interpretação

Com uma análise preliminar, observando apenas a quantidade de *snapshots* que não foram processados, aqui denominado de ganho, chega-se à conclusão que a melhor alternativa seja iniciar a análise o quanto antes, ou seja, com 20% dos *snapshots* processados. Ainda, ao se considerar que a função de similaridade é adequada à realidade em questão e que os dados estão corretamente agrupados e que o resultado de um seja, provavelmente, o resultado dos demais,

essa conclusão está correta. Apesar disso, outros estudos foram realizados, verificando se os *snapshots* que obtiveram os melhores resultados foram contemplados, ou seja, se foram processados quando as análises das diferentes situações foram feitas. A Tabela 16 apresenta os números dessa análise. A linha "*Melhores_10%*" refere-se aos 304 *snapshots* que, em um processamento exaustivo, apresentaram os melhores resultados de FEB. A linha "*Melhores_30%*" refere-se aos 912 *snapshots* que, em um processamento exaustivo, apresentaram os melhores resultados de FEB. As colunas indicam a cobertura desses *snapshots* quando as análises iniciaram com 20%, 30%, 50%, 70% e 80%, respectivamente, do processamento concluído.

Tabela 16 – *Snapshots* com melhores resultados relacionados à quantidade processada para análise

	Proc_20%	**Proc_30%**	**Proc_50%**	**Proc_70%**	**Proc_80%**
Melhores_10%	76%	79%	85%	89%	93%
Melhores_30%	70%	74%	82%	86%	90%

Ao se analisar a Tabela 16, verifica-se que os dados não estão agrupados de forma tão simétrica como se esperava, pois mesmo com a análise sendo realizada após 80% do total de *snapshots* terem sido processados, não se chega ao processamento de um número perto de 100% dos *snapshots* de melhor resultado. A variação entre os pontos de análise não é grande. Por mais que possa ser considerado que aproximadamente 24% dos *snapshots* com melhores resultados não estão sendo processados quando se inicia a análise em 20% e que aproximadamente 20% não são contemplados quando se inicia a análise com 30% dos *snapshots* processados, esse é um risco a ser assumido quando se deseja reduzir a quantidade de *snapshots* a serem processados.

A Figura 69 contém o gráfico com a análise dos resultados obtidos após a execução dos experimentos e a visualização da manutenção dos resultados promissores, que correspondam aos 10% melhores.

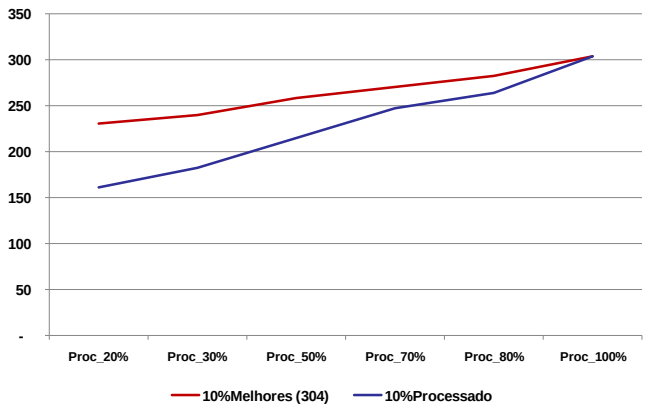

Figura 69 – Gráfico com análise do resultado, considerando 10% dos melhores resultados

No gráfico da Figura 69 verifica-se que o P-MIA supera a manutenção dos resultados promissores, quando comparado aos 10% dos *snapshots* processados. Da mesma forma, a Figura 70 contém o gráfico com a análise dos resultados obtidos após a execução dos experimentos e a visualização da manutenção dos resultados promissores, que correspondam aos 30% melhores.

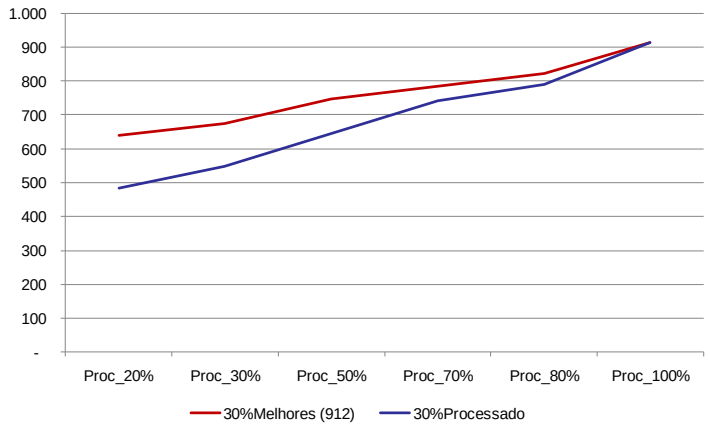

Figura 70 – Gráfico com análise do resultado, considerando 30% dos melhores resultados

7.3.3 Testes Experimentais com o Ligante NADH

Para os testes realizados com o ligante NADH, os valores foram especificados da mesma forma que para o ligante PIF, são eles:

- Quantidade mínima de *snapshots* para processamento por lote: 50
- Amostragem: 30%
- Melhor valor: -20,01
- Pior valor: -6,54

Tabela 17 - Resultado do Processamento de dois (2) *snapshots* de cada cluster para obter o melhor e pior valores - NADH

snapshot	***Cluster***	**FEB**
106	0	-18,22
46	0	-17,45
2748	1	-9,71
2766	1	-7,91
566	2	-7,83
412	2	-6,54
182	3	-20,01
248	3	-18,39
956	4	-16,59
867	4	-7,9
1467	5	-14,35
1574	5	-9,71

O intervalo gerado pelos valores fornecidos para [melhor valor, pior valor] é [-20,01; -6,54]. Esses valores foram obtidos por amostragem, executando-se 2 *snapshots* de cada grupo, de forma aleatória. Os resultados desse processamento são apresentados na Tabela 17.

7.3.3.1 Separação em Lotes

A Tabela 18 apresenta a quantidade de *snapshots* que compõem cada um dos *clusters*, gerados a partir do algoritmo *k-Means*.

Tabela 18 – Quantidade de *snapshots* em cada *cluster*, gerados pelo algoritmo k-means - NADH

Cluster	**Quantidade**
*Cluster*_0	144
*Cluster*_1	569
*Cluster*_2	508
*Cluster*_3	141
*Cluster*_4	544
*Cluster*_5	1194

Os lotes criados são encontrados no Apêndice A.

7.3.3.2 Resultados Obtidos

Os testes foram desenvolvidos da mesma forma que para o ligante PIF, já detalhados na seção 7.3.2. As Tabelas 19, 20, 21, 22 e 23 contêm os resultados dos testes realizados com o ligante NADH e, da mesma forma que para o PIF, o ganho de processamento foi considerável. O valor médio, utilizado como parâmetro para a realização destes testes é igual a -13,26, gerado a partir da média aritmética entre o melhor e pior valores: [-20,01; -6,54].

Nas Tabelas 19, 20, 21, 22 e 23, da mesma forma que para o ligante PIF, as células sombreadas e sem valor correspondem a *snapshots* que não foram processados após a análise; e as células sombreadas, com valor, correspondem a *snapshots* que foram processados antes do ponto de análise.

Na Na Tabela 20, com o ponto de análise iniciado aos 30% da totalidade de processamento, o ganho foi de 41%, ou seja, 1261 *snapshots* não foram processados.

Tabela 19, quando o ponto de análise foi iniciado com 20% do processamento concluído, o ganho total foi de 43%, totalizando 1332 de 3100 *snapshots* que não precisaram ser processados, considerando o valor obtido pela média aritmética dos resultados de FEB e pela média estimada.

Na Tabela 20, com o ponto de análise iniciado aos 30% da totalidade de processamento, o ganho foi de 41%, ou seja, 1261 *snapshots* não foram processados.

Tabela 19 – Análise dos resultados com 20% dos *snapshots* processados - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0_L0	50	-16,86	-18,70	10	-16,31	-18,02	15	-16,20	-17,51	25	-16,44	-17,21	35	-16,72	-17,21	40	50	0
C0_L1	94	-18,75	-19,84	19	-18,85	-19,57	28	-18,62	-19,04	47	-18,59	-18,85	66	-18,58	-18,67	75	94	0
C1_L0	171	-12,95	-14,23	34	-12,35	-13,63	51	-11,90	-12,91	86							86	85
C1_L1	120	-4,52	-15,99	24	-6,90	-15,14	36	-8,05	-12,63	60							60	60
C1_L2	83	-10,31	-11,74	17													17	66
C1_L3	58	-13,91	-14,80	12	-13,92	-14,56	17	-14,19	-14,65	29	-13,55	-14,00	41	-13,67	-13,81	46	58	0
C1_L4	50	-14,65	-15,70	10	-14,49	-15,51	15	-14,61	-15,32	25	-14,61	-15,02	35	-14,71	-14,96	40	50	0
C1_L5	87	-14,62	-15,40	17	-14,80	-15,63	26	-15,04	-15,74	44	-14,95	-15,37	61	-15,13	-15,45	70	87	0
C2_L0	152	50,80	-21,48	30	37,15	-18,03	46	30,98	-5,88	76							76	76
C2_L1	107	10,22	-2,67	21													21	86
C2_L2	75	2,75	-7,45	15													15	60
C2_L3	52	462,99	144,15	10													10	42
C2_L4	50	49,29	15,85	10													10	40
C2_L5	72	26,47	-14,80	14	63,98	10,85	22										22	50
C3_L0	50	-18,52	-18,71	10	-18,50	-18,64	15	-18,43	-18,78	25	19,79	-14,06	35	15,13	-5,97	40	50	0
C3_L1	91	-18,76	-19,04	18	-18,77	-18,94	27	-18,47	-19,05	46	-18,27	-18,61	64	-18,12	-18,36	73	91	0
C4_L0	163	9,22	-13,10	33													33	130
C4_L1	114	8,57	-28,69	23	25,84	-12,32	34										34	80
C4_L2	80	98,15	13,50	16													16	64
C4_L3	56	-13,22	-14,33	11	-13,68	-14,68	17	10,63	-15,19	28	3,67	-9,50	39				39	17
C4_L4	50	74,07	-39,87	10	44,13	-37,22	15	20,53	-24,44	25	10,42	-12,38	35				35	15
C4_L5	81	-13,83	-15,31	16	-13,77	-15,00	24	-11,55	-15,22	41	4,34	-14,55	57	1,86	-9,95	65	81	0
C5_L0	358	12,92	-59,11	72	3,96	-47,82	107	-3,23	-31,85	179	-5,40	-20,01	251	-6,33	-15,44	286	358	0
C5_L1	251	-12,28	-13,33	50	-12,08	-13,04	75										75	176
C5_L2	176	-13,73	-15,36	35	-13,16	-14,55	53	-12,75	-13,69	88	-12,45	-12,99	123				123	53
C5_L3	123	-12,66	-14,12	25	-12,48	-13,69	37	-11,90	-12,71	62							62	61
C5_L4	86	-11,24	-12,90	17													17	69
C5_L5	60	-12,71	-13,93	12	-11,65	-12,87	18										18	42
C5_L6	50	-13,31	-14,69	10	-12,99	-14,17	15	-13,00	-13,82	25	-12,66	-13,17	35				35	15
C5_L7	90	-12,93	-14,17	18	-12,77	-13,95	27	-11,38	-12,40	45							45	45
Ganho de Processamento – *snapshots*																		1332
Ganho de Processamento - percentual																		43%

Na Tabela 21, com o ponto de análise iniciado aos 50% da totalidade de processamento, o ganho foi de 33%, ou seja, 1021 *snapshots* não foram processados. Os percentuais de ganho, ao se testar o modelo com o ligante NADH, foram diferentes dos obtidos com os testes realizados com o ligante PIF. Apesar disso, a diferença não se apresentou muito grande, sendo de aproximadamente 3% para mais ou para menos nos diferentes pontos de análise.

Tabela 20 – Análise dos resultados com 30% dos *snapshots* processados - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0_L0	50	-16,86	-18,70	10	-16,31	-18,02	15	-16,20	-17,51	25	-16,44	-17,21	35	-16,72	-17,21	40	50	0
C0_L1	94	-18,75	-19,84	19	-18,85	-19,57	28	-18,62	-19,04	47	-18,59	-18,85	66	-18,58	-18,67	75	94	0
C1_L0	171	-12,95	-14,23	34	-12,35	-13,63	51	-11,90	-12,91	86							86	85
C1_L1	120	-4,52	-15,99	24	-6,90	-15,14	36	-8,05	-12,63	60							60	60
C1_L2	83	-10,31	-11,74	17	-10,14	-11,29	25										25	58
C1_L3	58	-13,91	-14,80	12	-13,92	-14,56	17	-14,19	-14,65	29	-13,55	-14,00	41	-13,67	-13,81	46	58	0
C1_L4	50	-14,65	-15,70	10	-14,49	-15,51	15	-14,61	-15,32	25	-14,61	-15,02	35	-14,71	-14,96	40	50	0
C1_L5	87	-14,62	-15,40	17	-14,80	-15,63	26	-15,04	-15,74	44	-14,95	-15,37	61	-15,13	-15,45	70	87	0
C2_L0	152	50,80	-21,48	30	37,15	-18,03	46	30,98	-5,88	76							76	76
C2_L1	107	10,22	-2,67	21	3,93	-5,74	32										32	75
C2_L2	75	2,75	-7,45	15	16,76	-4,45	23										23	52
C2_L3	52	462,99	144,15	10	320,41	96,96	16										16	36
C2_L4	50	49,29	15,85	10	86,82	14,81	15										15	35
C2_L5	72	26,47	-14,80	14	63,98	10,85	22										22	50
C3_L0	50	-18,52	-18,71	10	-18,50	-18,64	15	-18,43	-18,78	25	19,79	-14,06	35	15,13	-5,97	40	50	0
C3_L1	91	-18,76	-19,04	18	-18,77	-18,94	27	-18,47	-19,05	46	-18,27	-18,61	64	-18,12	-18,36	73	91	0
C4_L0	163	9,22	-13,10	33	23,43	-9,81	49										49	114
C4_L1	114	8,57	-28,69	23	25,84	-12,32	34										34	80
C4_L2	80	98,15	13,50	16	148,73	-8,74	24										24	56
C4_L3	56	-13,22	-14,33	11	-13,68	-14,68	17	10,63	-15,19	28	3,67	-9,50	39				39	17
C4_L4	50	74,07	-39,87	10	44,13	-37,22	15	20,53	-24,44	25	10,42	-12,38	35				35	15
C4_L5	81	-13,83	-15,31	16	-13,77	-15,00	24	-11,55	-15,22	41	4,34	-14,55	57	1,86	-9,95	65	81	0
C5_L0	358	12,92	-59,11	72	3,96	-47,82	107	-3,23	-31,85	179	-5,40	-20,01	251	-6,33	-15,44	286	358	0
C5_L1	251	-12,28	-13,33	50	-12,08	-13,04	75										75	176
C5_L2	176	-13,73	-15,36	35	-13,16	-14,55	53	-12,75	-13,69	88	-12,45	-12,99	123				123	53
C5_L3	123	-12,66	-14,12	25	-12,48	-13,69	37	-11,90	-12,71	62							62	61
C5_L4	86	-11,24	-12,90	17	-10,55	-11,95	26										26	60
C5_L5	60	-12,71	-13,93	12	-11,65	-12,87	18										18	42
C5_L6	50	-13,31	-14,69	10	-12,99	-14,17	15	-13,00	-13,82	25	-12,66	-13,17	35				35	15
C5_L7	90	-12,93	-14,17	18	-12,77	-13,95	27	-11,38	-12,40	45							45	45
Ganho de Processamento – *snapshots*																		1261
Ganho de Processamento - percentual																		41%

Na Tabela 22, o ponto de análise iniciou após 70% do total de *snapshots* de cada lote estar concluído. Mesmo assim, obtém-se um ganho satisfatório, ao se considerar que a partir da média aritmética e da média estimada, aproximadamente 21% dos *snapshots*, ou seja, 655, não foram processados. Para o ligante PIF, apresentado na seção anterior, esse ganho ficou em 19%.

Na Tabela 23, quando o ponto de análise foi fixado para iniciar após 80% do processamento estar concluído, o ganho foi de 14%: 435 *snapshots* não foram processados.

Tabela 21 – Análise dos resultados com 50% dos *snapshots* processados - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0_L0	50	-16,86	-18,70	10	-16,31	-18,02	15	-16,20	-17,51	25	-16,44	-17,21	35	-16,72	-17,21	40	50	0
C0_L1	94	-18,75	-19,84	19	-18,85	-19,57	28	-18,62	-19,04	47	-18,59	-18,85	66	-18,58	-18,67	75	94	0
C1_L0	171	-12,95	-14,23	34	-12,35	-13,63	51	-11,90	-12,91	86							86	85
C1_L1	120	-4,52	-15,99	24	-6,90	-15,14	36	-8,05	-12,63	60							60	60
C1_L2	83	-10,31	-11,74	17	-10,14	-11,29	25	-9,67	-10,46	42							42	41
C1_L3	58	-13,91	-14,80	12	-13,92	-14,56	17	-14,19	-14,65	29	-13,55	-14,00	41	-13,67	-13,81	46	58	0
C1_L4	50	-14,65	-15,70	10	-14,49	-15,51	15	-14,61	-15,32	25	-14,61	-15,02	35	-14,71	-14,96	40	50	0
C1_L5	87	-14,62	-15,40	17	-14,80	-15,63	26	-15,04	-15,74	44	-14,95	-15,37	61	-15,13	-15,45	70	87	0
C2_L0	152	50,80	-21,48	30	37,15	-18,03	46	30,98	-5,88	76							76	76
C2_L1	107	10,22	-2,67	21	3,93	-5,74	32	5,85	-0,35	54							54	53
C2_L2	75	2,75	-7,45	15	16,76	-4,45	23	43,95	18,79	38							38	37
C2_L3	52	462,99	144,15	10	320,41	96,96	16	214,83	85,14	26							26	26
C2_L4	50	49,29	15,85	10	86,82	14,81	15	53,26	12,47	25							25	25
C2_L5	72	26,47	-14,80	14	63,98	10,85	22	49,06	17,46	36							36	36
C3_L0	50	-18,52	-18,71	10	-18,50	-18,64	15	-18,43	-18,78	25	19,79	-14,06	35	15,13	-5,97	40	50	0
C3_L1	91	-18,76	-19,04	18	-18,77	-18,94	27	-18,47	-19,05	46	-18,27	-18,61	64	-18,12	-18,36	73	91	0
C4_L0	163	9,22	-13,10	33	23,43	-9,81	49	59,92	6,99	82							82	81
C4_L1	114	8,57	-28,69	23	25,84	-12,32	34	10,43	-10,99	57							57	57
C4_L2	80	98,15	13,50	16	148,73	-8,74	24	92,17	3,44	40							40	40
C4_L3	56	-13,22	-14,33	11	-13,68	-14,68	17	10,63	-15,19	28	3,67	-9,50	39				39	17
C4_L4	50	74,07	-39,87	10	44,13	-37,22	15	20,53	-24,44	25	10,42	-12,38	35				35	15
C4_L5	81	-13,83	-15,31	16	-13,77	-15,00	24	-11,55	-15,22	41	4,34	-14,55	57	1,86	-9,95	65	81	0
C5_L0	358	12,92	-59,11	72	3,96	-47,82	107	-3,23	-31,85	179	-5,40	-20,01	251	-6,33	-15,44	286	358	0
C5_L1	251	-12,28	-13,33	50	-12,08	-13,04	75	-12,12	-12,86	126							126	125
C5_L2	176	-13,73	-15,36	35	-13,16	-14,55	53	-12,75	-13,69	88	-12,45	-12,99	123				123	53
C5_L3	123	-12,66	-14,12	25	-12,48	-13,69	37	-11,90	-12,71	62							62	61
C5_L4	86	-11,24	-12,90	17	-10,55	-11,95	26	-10,86	-11,77	43							43	43
C5_L5	60	-12,71	-13,93	12	-11,65	-12,87	18	-11,21	-11,96	30							30	30
C5_L6	50	-13,31	-14,69	10	-12,99	-14,17	15	-13,00	-13,82	25	-12,66	-13,17	35				35	15
C5_L7	90	-12,93	-14,17	18	-12,77	-13,95	27	-11,38	-12,40	45							45	45
Ganho de Processamento – *snapshots*																		1021
Ganho de Processamento - percentual																		33%

Tabela 22 – Análise dos resultados com 70% dos *snapshots* processados - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0_L0	50	-16,86	-18,70	10	-16,31	-18,02	15	-16,20	-17,51	25	-16,44	-17,21	35	-16,72	-17,21	40	50	0
C0_L1	94	-18,75	-19,84	19	-18,85	-19,57	28	-18,62	-19,04	47	-18,59	-18,85	66	-18,58	-18,67	75	94	0
C1_L0	171	-12,95	-14,23	34	-12,35	-13,63	51	-11,90	-12,91	86	-11,26	-11,89	120				120	51
C1_L1	120	-4,52	-15,99	24	-6,90	-15,14	36	-8,05	-12,63	60	-8,32	-10,64	84				84	36
C1_L2	83	-10,31	-11,74	17	-10,14	-11,29	25	-9,67	-10,46	42	-10,45	-10,92	58				58	25
C1_L3	58	-13,91	-14,80	12	-13,92	-14,56	17	-14,19	-14,65	29	-13,55	-14,00	41	-13,67	-13,81	46	58	0
C1_L4	50	-14,65	-15,70	10	-14,49	-15,51	15	-14,61	-15,32	25	-14,61	-15,02	35	-14,71	-14,96	40	50	0
C1_L5	87	-14,62	-15,40	17	-14,80	-15,63	26	-15,04	-15,74	44	-14,95	-15,37	61	-15,13	-15,45	70	87	0
C2_L0	152	50,80	-21,48	30	37,15	-18,03	46	30,98	-5,88	76	23,15	3,64	106				106	46
C2_L1	107	10,22	-2,67	21	3,93	-5,74	32	5,85	-0,35	54	11,44	6,74	75				75	32
C2_L2	75	2,75	-7,45	15	16,76	-4,45	23	43,95	18,79	38	51,66	36,89	53				53	22
C2_L3	52	462,99	144,15	10	320,41	96,96	16	214,83	85,14	26	223,47	152,35	36				36	16
C2_L4	50	49,29	15,85	10	86,82	14,81	15	53,26	12,47	25	37,00	15,95	35				35	15
C2_L5	72	26,47	-14,80	14	63,98	10,85	22	49,06	17,46	36	39,16	22,28	50				50	22
C3_L0	50	-18,52	-18,71	10	-18,50	-18,64	15	-18,43	-18,78	25	19,79	-14,06	35	15,13	-5,97	40	50	0
C3_L1	91	-18,76	-19,04	18	-18,77	-18,94	27	-18,47	-19,05	46	-18,27	-18,61	64	-18,12	-18,36	73	91	0
C4_L0	163	9,22	-13,10	33	23,43	-9,81	49	59,92	6,99	82	116,42	55,97	114				114	49
C4_L1	114	8,57	-28,69	23	25,84	-12,32	34	10,43	-10,99	57	13,14	-0,09	80				80	34
C4_L2	80	98,15	13,50	16	148,73	-8,74	24	92,17	3,44	40	85,88	39,21	56				56	24
C4_L3	56	-13,22	-14,33	11	-13,68	-14,68	17	10,63	-15,19	28	3,67	-9,50	39				39	17
C4_L4	50	74,07	-39,87	10	44,13	-37,22	15	20,53	-24,44	25	10,42	-12,38	35				35	15
C4_L5	81	-13,83	-15,31	16	-13,77	-15,00	24	-11,55	-15,22	41	4,34	-14,55	57	1,86	-9,95	65	81	0
C5_L0	358	12,92	-59,11	72	3,96	-47,82	107	-3,23	-31,85	179	-5,40	-20,01	251	-6,33	-15,44	286	358	0
C5_L1	251	-12,28	-13,33	50	-12,08	-13,04	75	-12,12	-12,86	126	-12,01	-12,46	176				176	75
C5_L2	176	-13,73	-15,36	35	-13,16	-14,55	53	-12,75	-13,69	88	-12,45	-12,99	123				123	53
C5_L3	123	-12,66	-14,12	25	-12,48	-13,69	37	-11,90	-12,71	62	-11,77	-12,22	86				86	37
C5_L4	86	-11,24	-12,90	17	-10,55	-11,95	26	-10,86	-11,77	43	-11,42	-11,93	60				60	26
C5_L5	60	-12,71	-13,93	12	-11,65	-12,87	18	-11,21	-11,96	30	-11,74	-12,21	42				42	18
C5_L6	50	-13,31	-14,69	10	-12,99	-14,17	15	-13,00	-13,82	25	-12,66	-13,17	35				35	15
C5_L7	90	-12,93	-14,17	18	-12,77	-13,95	27	-11,38	-12,40	45	-10,76	-11,37	63				63	27
Ganho de Processamento – *snapshots*																		655
Ganho de Processamento - percentual																		21%

Tabela 23 – Análise dos resultados com 80% dos *snapshots* processados - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0_L0	50	-16,86	-18,70	10	-16,31	-18,02	15	-16,20	-17,51	25	-16,44	-17,21	35	-16,72	-17,21	40	50	0
C0_L1	94	-18,75	-19,84	19	-18,85	-19,57	28	-18,62	-19,04	47	-18,59	-18,85	66	-18,58	-18,67	75	94	0
C1_L0	171	-12,95	-14,23	34	-12,35	-13,63	51	-11,90	-12,91	86	-11,26	-11,89	120	-11,20	-11,63	137	137	34
C1_L1	120	-4,52	-15,99	24	-6,90	-15,14	36	-8,05	-12,63	60	-8,32	-10,64	84	-8,55	-10,00	96	96	24
C1_L2	83	-10,31	-11,74	17	-10,14	-11,29	25	-9,67	-10,46	42	-10,45	-10,92	58	-10,73	-11,00	66	66	17
C1_L3	58	-13,91	-14,80	12	-13,92	-14,56	17	-14,19	-14,65	29	-13,55	-14,00	41	-13,67	-13,81	46	58	0
C1_L4	50	-14,65	-15,70	10	-14,49	-15,51	15	-14,61	-15,32	25	-14,61	-15,02	35	-14,71	-14,96	40	50	0
C1_L5	87	-14,62	-15,40	17	-14,80	-15,63	26	-15,04	-15,74	44	-14,95	-15,37	61	-15,13	-15,45	70	87	0
C2_L0	152	50,80	-21,48	30	37,15	-18,03	46	30,98	-5,88	76	23,15	3,64	106	20,06	7,93	122	122	30
C2_L1	107	10,22	-2,67	21	3,93	-5,74	32	5,85	-0,35	54	11,44	6,74	75	14,54	10,29	86	86	21
C2_L2	75	2,75	-7,45	15	16,76	-4,45	23	43,95	18,79	38	51,66	36,89	53	53,37	42,95	60	60	15
C2_L3	52	462,99	144,15	10	320,41	96,96	16	214,83	85,14	26	223,47	152,35	36	210,84	169,19	42	42	10
C2_L4	50	49,29	15,85	10	86,82	14,81	15	53,26	12,47	25	37,00	15,95	35	31,16	17,97	40	40	10
C2_L5	72	26,47	-14,80	14	63,98	10,85	22	49,06	17,46	36	39,16	22,28	50	35,53	25,29	58	58	14
C3_L0	50	-18,52	-18,71	10	-18,50	-18,64	15	-18,43	-18,78	25	19,79	-14,06	35	15,13	-5,97	40	50	0
C3_L1	91	-18,76	-19,04	18	-18,77	-18,94	27	-18,47	-19,05	46	-18,27	-18,61	64	-18,12	-18,36	73	91	0
C4_L0	163	9,22	-13,10	33	23,43	-9,81	49	59,92	6,99	82	116,42	55,97	114	113,47	72,94	130	130	33
C4_L1	114	8,57	-28,69	23	25,84	-12,32	34	10,43	-10,99	57	13,14	-0,09	80	57,80	36,12	91	91	23
C4_L2	80	98,15	13,50	16	148,73	-8,74	24	92,17	3,44	40	85,88	39,21	56	75,01	45,78	64	64	16
C4_L3	56	-13,22	-14,33	11	-13,68	-14,68	17	10,63	-15,19	28	3,67	-9,50	39	1,34	-6,83	45	45	11
C4_L4	50	74,07	-39,87	10	44,13	-37,22	15	20,53	-24,44	25	10,42	-12,38	35	7,21	-7,00	40	40	10
C4_L5	81	-13,83	-15,31	16	-13,77	-15,00	24	-11,55	-15,22	41	4,34	-14,55	57	1,86	-9,95	65	81	0
C5_L0	358	12,92	-59,11	72	3,96	-47,82	107	-3,23	-31,85	179	-5,40	-20,01	251	-6,33	-15,44	286	358	0
C5_L1	251	-12,28	-13,33	50	-12,08	-13,04	75	-12,12	-12,86	126	-12,01	-12,46	176	-12,14	-12,44	201	201	50
C5_L2	176	-13,73	-15,36	35	-13,16	-14,55	53	-12,75	-13,69	88	-12,45	-12,99	123	-12,49	-12,88	141	141	35
C5_L3	123	-12,66	-14,12	25	-12,48	-13,69	37	-11,90	-12,71	62	-11,77	-12,22	86	-11,43	-11,71	98	98	25
C5_L4	86	-11,24	-12,90	17	-10,55	-11,95	26	-10,86	-11,77	43	-11,42	-11,93	60	-11,50	-11,89	69	69	17
C5_L5	60	-12,71	-13,93	12	-11,65	-12,87	18	-11,21	-11,96	30	-11,74	-12,21	42	-11,83	-12,14	48	48	12
C5_L6	50	-13,31	-14,69	10	-12,99	-14,17	15	-13,00	-13,82	25	-12,66	-13,17	35	-12,89	-13,23	40	40	10
C5_L7	90	-12,93	-14,17	18	-12,77	-13,95	27	-11,38	-12,40	45	-10,76	-11,37	63	-10,91	-11,34	72	72	18
Ganho de Processamento – *snapshots*																		435
Ganho de Processamento - percentual																		14%

7.3.3.3 Análise e Interpretação

As Tabelas 24, 25, 26, 27 e 28 contêm os testes realizados para o ligante NADH, considerando o grupo como um todo, sem a separação em lotes menores. Verifica-se, a partir destes testes, que o ganho em processamento é bem menor e não apresenta grande variação caso as análises sejam realizadas mais cedo ou mais tarde, ou seja, não importa se a análise é iniciada com 20% (Tabela 24), 30% (Tabela 25) ou 50% (Tabela 26), pois em todos esses casos o ganho de processamento foi o mesmo, apenas de 22%. Esse caso é diferente da análise realizada por meio de lotes menores que, com 20%, obteve-se um ganho de 43%, ou seja, 43% dos *snapshots* não foram processados; com 30% obteve-se um ganho de 41%; e com 50% dos *snapshots* processados, obteve-se um ganho de 33%. Com 70% (Tabela 27) e com 80% (Tabela 28) do processamento

concluído para se dar início às análises, os valores são ainda piores: 16% e 10% de ganho respectivamente.

Tabela 24 – Análise dos resultados com 20% dos *snapshots* processados sem separação em lotes - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0	144	-15,80	-18,01	29	-16,92	-18,55	43	-16,12	-19,51	72	-16,79	-18,54	101	-17,01	-18,06	115	144	0
C1	569	-11,00	-13,29	114	-7,25	-17,72	171	-7,82	-13,99	285	-8,94	-12,08	398				398	171
C2	508	24,47	-28,33	102	14,10	-24,31	152	13,54	-8,62	254							254	254
C3	141	-18,43	-18,95	28	13,64	-58,47	42	0,36	-39,23	71	-4,85	-24,98	99	-6,48	-19,03	113	141	0
C4	544	122,32	-41,94	109	88,64	-38,32	163	86,17	7,49	272							272	272
C5	1194	-5,00	-44,90	239	-7,63	-36,17	358	-9,56	-25,35	597	-10,43	-18,44	836	-10,41	-15,41	955	1194	0
Ganho de Processamento – *snapshots*																		697
Ganho de Processamento - percentual																		22%

Tabela 25 – Análise dos resultados com 30% dos *snapshots* processados sem separação em lotes - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0	144	-15,80	-18,01	29	-16,92	-18,55	43	-16,12	-19,51	72	-16,79	-18,54	101	-17,01	-18,06	115	144	0
C1	569	-11,00	-13,29	114	-7,25	-17,72	171	-7,82	-13,99	285	-8,94	-12,08	398				398	171
C2	508	24,47	-28,33	102	14,10	-24,31	152	13,54	-8,62	254							254	254
C3	141	-18,43	-18,95	28	13,64	-58,47	42	0,36	-39,23	71	-4,85	-24,98	99	-6,48	-19,03	113	141	0
C4	544	122,32	-41,94	109	88,64	-38,32	163	86,17	7,49	272							272	272
C5	1194	-5,00	-44,90	239	-7,63	-36,17	358	-9,56	-25,35	597	-10,43	-18,44	836	-10,41	-15,41	955	1194	0
Ganho de Processamento – *snapshots*																		697
Ganho de Processamento - percentual																		22%

Tabela 26 – Análise dos resultados com 50% dos *snapshots* processados sem separação em lotes - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0	144	-15,80	-18,01	29	-16,92	-18,55	43	-16,12	-19,51	72	-16,79	-18,54	101	-17,01	-18,06	115	144	0
C1	569	-11,00	-13,29	114	-7,25	-17,72	171	-7,82	-13,99	285	-8,94	-12,08	398				398	171
C2	508	24,47	-28,33	102	14,10	-24,31	152	13,54	-8,62	254							254	254
C3	141	-18,43	-18,95	28	13,64	-58,47	42	0,36	-39,23	71	-4,85	-24,98	99	-6,48	-19,03	113	141	0
C4	544	122,32	-41,94	109	88,64	-38,32	163	86,17	7,49	272							272	272
C5	1194	-5,00	-44,90	239	-7,63	-36,17	358	-9,56	-25,35	597	-10,43	-18,44	836	-10,41	-15,41	955	1194	0
Ganho de Processamento – *snapshots*																		697
Ganho de Processamento - percentual																		22%

Tabela 27 – Análise dos resultados com 70% dos *snapshots* processados sem separação em lotes - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0	144	-15,80	-18,01	29	-16,92	-18,55	43	-16,12	-19,51	72	-16,79	-18,54	101	-17,01	-18,06	115	144	0
C1	569	-11,00	-13,29	114	-7,25	-17,72	171	-7,82	-13,99	285	-8,94	-12,08	398				398	171
C2	508	24,47	-28,33	102	14,10	-24,31	152	13,54	-8,62	254	43,92	14,04	356				356	152
C3	141	-18,43	-18,95	28	13,64	-58,47	42	0,36	-39,23	71	-4,85	-24,98	99	-6,48	-19,03	113	141	0
C4	544	122,32	-41,94	109	88,64	-38,32	163	86,17	7,49	272	77,33	32,50	381				381	163
C5	1194	-5,00	-44,90	239	-7,63	-36,17	358	-9,56	-25,35	597	-10,43	-18,44	836	-10,41	-15,41	955	1194	0
Ganho de Processamento – *snapshots*																		486
Ganho de Processamento - percentual																		16%

Tabela 28 – Análise dos resultados com 80% dos *snapshots* processados sem separação em lotes - NADH

C_L	Quant	M20%	E20%	Proc	M30%	E30%	Proc	M50%	E50%	Proc	M70%	E70%	Proc	M80%	E80%	Proc	ProcFinal	Ganho
C0	144	-15,80	-18,01	29	-16,92	-18,55	43	-16,12	-19,51	72	-16,79	-18,54	101	-17,01	-18,06	115	144	0
C1	569	-11,00	-13,29	114	-7,25	-17,72	171	-7,82	-13,99	285	-8,94	-12,08	398	-9,55	-11,52	455	455	114
C2	508	24,47	-28,33	102	14,10	-24,31	152	13,54	-8,62	254	43,92	14,04	356	52,79	32,48	406	406	102
C3	141	-18,43	-18,95	28	13,64	-58,47	42	0,36	-39,23	71	-4,85	-24,98	99	-6,48	-19,03	113	141	0
C4	544	122,32	-41,94	109	88,64	-38,32	163	86,17	7,49	272	77,33	32,50	381	69,47	40,95	435	435	109
C5	1194	-5,00	-44,90	239	-7,63	-36,17	358	-9,56	-25,35	597	-10,43	-18,44	836	-10,41	-15,41	955	1194	0
Ganho de Processamento – *snapshots*																		325
Ganho de Processamento - percentual																		10%

Tabela 29 – *Snapshots* com melhores resultados relacionados à quantidade processada para análise - NADH

	Proc_20%	Proc_30%	Proc_50%	Proc_70%	Proc_80%
Melhores_10%	89%	89%	90%	92%	94%
Melhores_30%	73%	74%	77%	82%	88%

Ao se analisar a Tabela 29, da mesma forma que para o ligante PIF, verifica-se que os dados não estão agrupados de forma tão simétrica como se esperava, pois mesmo com a análise sendo realizada após 80% do total de *snapshots* terem sido processados, não se chega ao processamento de um número perto de 100% dos *snapshots* de melhor resultado. O desvio padrão dos resultados de FEB, obtidos após o processamento dos diferentes grupos, é muito inconstante, chegando, em alguns lotes, a valor igual a 114, por exemplo, enquanto outros lotes apresentam desvio padrão igual a 0,91 o que demonstra que o critério de similaridade utilizado não está fornecendo uma constância aos dados, já que lotes de um mesmo grupo apresentam valores muito distintos. Na Tabela 29 também se verifica que a análise sendo iniciada com 20% do processamento concluído, ou com 30%, obteve a mesma quantidade de *snapshots* com melhor resultado. Dessa forma, como o ganho de processamento é muito maior quanto mais cedo se iniciam as análises, sugere-se que o início das análises do P-MIA seja realizado, de forma pré-definida, quando 20% do processamento dos *snapshots* estiver concluído, sendo que esse valor pode ser reduzido ou incrementado antes do início do funcionamento do padrão.

Tabela 30 – *Snapshots* com melhores resultados relacionados à quantidade processada para análise, sem separação em lotes - NADH

	Proc_20%	Proc_30%	Proc_50%	Proc_70%	Proc_80%
Melhores_10%	92%	92%	92%	93%	94%
Melhores_30%	81%	81%	81%	83%	87%

A Tabela 30 contém a quantidade de *snapshots* que apresentaram os melhores resultados de FEB, que efetivamente foram processados por meio da aplicação do P-MIA, sem a separação dos grupos em lotes menores. Nota-se que a quantidade de *snapshots* processados foi um pouco maior que a separação em lotes, 3%, considerando-se 20%, 30% e 50% do processamento. Apesar disso, como o ganho em processamento, ao se separar em lotes menores, é muito superior, justifica-se a criação de lotes menores do P-MIA.

A Figura 71 contém os gráficos que apresentam a manutenção dos resultados mais promissores, considerando os 10% melhores, quando o experimento foi realizado separando-se os grupos em lotes menores e sem a separação. Nota-se o ganho obtido no processamento de lotes menores.

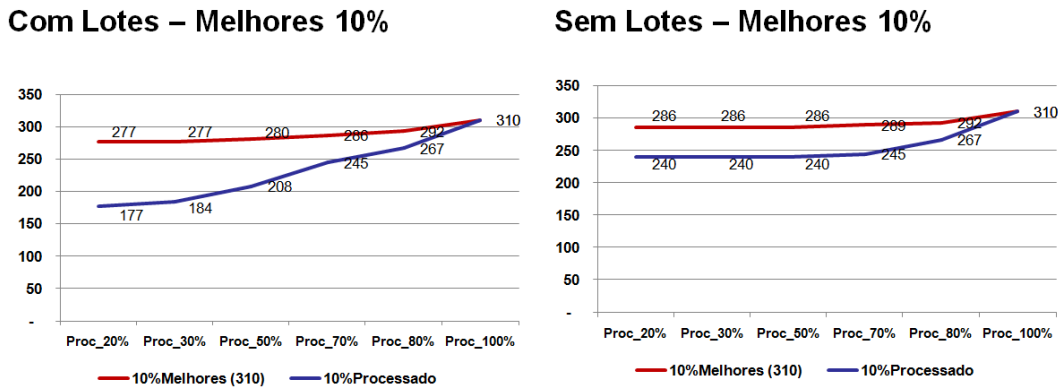

Figura 71 – Gráfico com análise do resultado, considerando manutenção dos valores com o processamento em grupos e em lotes

7.4 Considerações do Capítulo

Os testes realizados neste capítulo buscam a validação do Padrão Múltiplas Instâncias Autoadaptáveis (P-MIA), formalizado no Capítulo 5 e cujo funcionamento foi apresentado no Capítulo 6, subsidiando a definição das regras do padrão. Com a realização dos testes experimentais a hipótese nula definida como *"A utilização do P-MIA não resulta em ganhos"* foi negada, pois, os resultados obtidos com o experimento realizado com o ligante NADH, sintetizados na Tabela 31, ao se manipular lotes de *snapshots* quando, por exemplo, 20% do processamento está concluído, comprova essa afirmação. Nessa tabela, considerando-se uma quantidade total de *snapshots* a serem processados igual a 3100 e que desses, 1332 não foram processados quando a análise foi iniciada aos 20% do processamento, conferindo um ganho de 43% ao experimento, a expectativa poderia ser de que 43% dos *snapshots* de melhor resultado também não seriam processados. Esse número ficou em apenas 11%, pois 89% dos *snapshots* de melhor resultado foram contemplados quando a análise foi iniciada aos 20%.

Tabela 31 – Sintetização dos Resultados

Total *Snapshots*	10% dos Melhores Resultados Proc (20%)	Quantidade *Snapshot* Melhores resultados Contemplados	Ganho Proc (20%)	Quantidade *Snapshots* que não foram processados
3100	89%	276	43%	1332

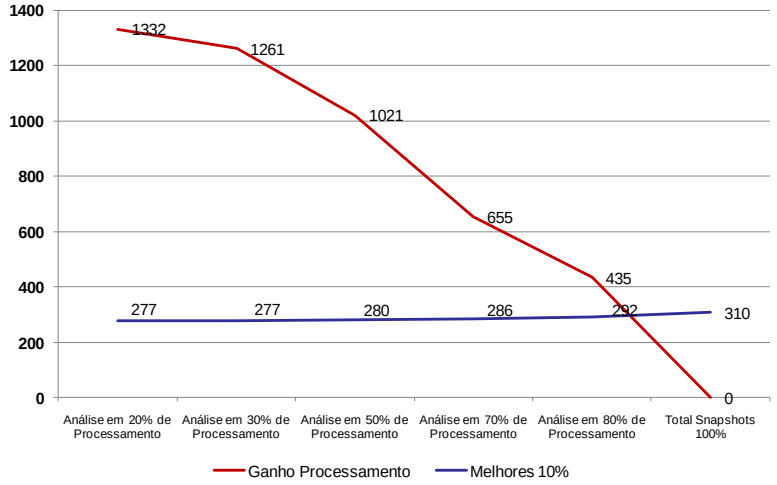

Figura 72 – Gráfico com análise do resultado e do ganho obtido – ligante NADH

Na Figura 72 visualiza-se o ganho obtido quanto antes as análises são realizadas e a manutenção dos resultados mais promissores, considerando os 10% melhores com o ligante NADH. A análise para o ligante PIF pode ser visualizada na Figura 73.

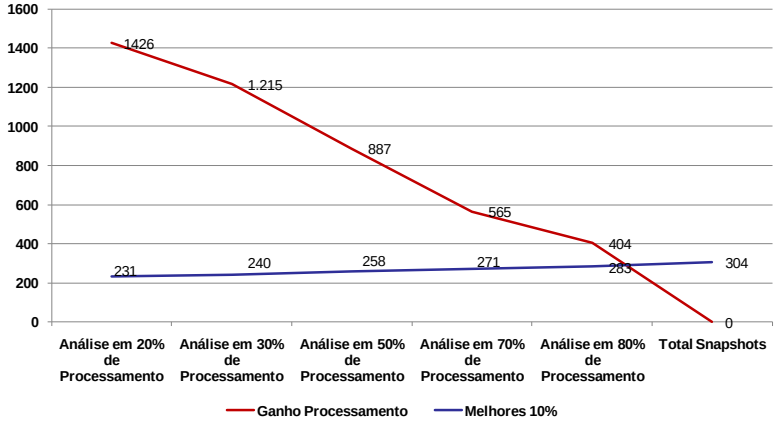

Figura 73 – Gráfico com análise do resultado e do ganho obtido – ligante PIF

Assim, os testes realizados demonstraram os ganhos obtidos com a aplicação do padrão, considerados como a redução da quantidade total de execuções e a continuidade de processamento dos *snapshots* que, em um processo exaustivo, apresentariam os melhores resultados. Dessa forma, a hipótese alternativa definida como *"A utilização do P-MIA resulta em ganhos"* foi comprovada. Além da validação das hipóteses, os testes realizados validaram as características do padrão: criação de lotes, aplicação da média aritmética e da média estimada,

descarte de *snapshots* e alteração de prioridade. Também foi possível, a partir da realização dos testes, que as seguintes questões obtivessem respostas:

- Grupos menores produzem resultados melhores que grupos maiores? SIM. Conforme apresentado pelos testes realizados neste capítulo, o processamento de grupos menores de *snapshots* gera maior ganho, pois as análises são realizadas com uma população menor.
- Quanto antes for iniciada a análise, maior é o ganho de processamento? SIM, pois se inicia o descarte de *snapshots*, de lotes que não apresentam bons resultados, antes, fazendo com que uma quantidade muito maior de *snapshots* não necessite de processamento.
- A função de similaridade aplicada está diretamente relacionada ao ganho obtido após a execução do padrão? SIM. Acredita-se que o trabalho que está em desenvolvimento por Karina Machado para a definição de uma função de similaridade específica para a área de estudo do LABIO, sobre os mesmos dados manipulados nesta Tese, melhore ainda mais os resultados obtidos, reduzindo a variação dos desvios padrão em um mesmo grupo, pois tende a criar grupos cujos resultados de FEB sejam mais próximos e, quando esses *snapshots* forem submetidos aos P-MIA, que o padrão possa obter uma maior cobertura dos melhores resultados.

Conclui-se, ainda, que os testes apresentados neste capítulo foram fundamentais para a validação do padrão e para a justificativa dos critérios definidos, podendo ser facilmente reproduzidos por meio da definição dos lotes e da implementação das funções de média e média estimada, apresentadas no Capítulo 6.

8 TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados ao desenvolvido nesta Tese, identificando semelhanças e diferenças. Cabe destacar que são apresentados onze trabalhos que apresentam características que, de alguma forma, relacionam-se com o trabalho desenvolvido, pois o foco da análise está na identificação de iniciativas, envolvendo: utilização de *workflows* científicos, principalmente na área de Bioinformática; formalismos de representação de processos, em especial redes de Petri; e utilização de Padrões de Dados, com o intuito de estabelecer pontos de interseção entre a Tese desenvolvida e os demais trabalhos da área.

8.1 Fluxos de Dados e Validações em Modelagens de *Workflows*

Sadiq et al., [SAD04], definem que a especificação completa de um *workflow* requer a integração de diferentes características de processos, como decisões, definições de atividades individuais, lógica do processo e regras de execução. Afirmam que uma das importantes características de modelagem e especificação de *workflows* envolve o fluxo dos dados, sua modelagem, especificação e validação e que muitos pesquisadores têm negligenciado esta dimensão da análise de processos. Os autores identificam e justificam a importância da modelagem de dados na especificação e verificação de *workflows*. Identificam possibilidades de problemas com fluxos de dados que, caso não sejam identificados, podem prejudicar o funcionamento dos *workflows*.

Sadiq et al., [SAD04], também afirmam que a tecnologia de *workflows* é utilizada como uma tecnologia de integração de sistemas existentes, uma prática muito utilizada em Bioinformática, aplicação foco desta Tese. Os autores apresentam a distinção clara entre aplicações específicas de dados em atividades individuais e de dados relacionados a controle de processos. Além disso, distinguem dados de entrada dos de saída de uma atividade. Essa distinção também é aplicada na Tese desenvolvida, sendo utilizada na definição do P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis.

Os autores em [SAD04] apresentam a diferenciação entre fontes de dados internas e fontes externas. Essa característica também é utilizada pelo P-MIA, uma vez que se trabalha com dados manipulados internamente no processo, onde tarefas geram valores e esses são utilizados por outras tarefas no processo (fonte de dados interna) e estruturas externas como bancos de dados, sendo acessíveis por tarefas do processo (fonte de dados externa). Apesar de apresentar vários conceitos que se relacionam com o trabalho desenvolvido nesta Tese, o trabalho de Sadiq et al. [SAD04] possui foco principal na validação da modelagem dos fluxos de dados, enquanto o desenvolvido apresenta um novo padrão de dados, cujas características envolvem a manipulação de grandes conjuntos de dados e a análise da continuidade do seu processamento.

8.2 Apoio a Fluxos de Dados Avançados para Aplicações de *Workflows* Científicos em Grade

Qin e Fahringer [QIN07] afirmam que a definição de estruturas de fluxo de dados direcionada à execução de conjuntos flexíveis de dados é um dos requisitos de execução para aplicações científicas em grade. Apresentam uma alternativa para a introdução do conceito de coleção de dados e sua correspondente distribuição em aplicações de *workflows* em grade. Afirmam, também, que a tecnologia em grade fez com que cientistas e engenheiros criassem aplicações cada vez mais complexas, para gerenciar grandes volumes de conjuntos de dados, e executar experimentos científicos sobre essa tecnologia.

Os autores definem que uma aplicação de *workflow* em grade pode ser vista como uma coleção de tarefas computacionais que são processadas em uma ordem bem definida, para atingir a um objetivo específico. Afirmam que diferentes construtores de controle de fluxo têm sido identificados e desenvolvidos para sistemas de *workflows* em grade, podendo ser divididos em quatro categorias: sequencial, paralela, condicional e interativa. Com cada um desses construtores, diferentes fluxos de dados podem ser especificados. Ainda destacam que fluxos de dados manipulados sobre *workflows* científicos em grade são geralmente complexos pelo conjunto de dados envolvidos e essa é uma das motivações do desenvolvimento desta Tese: a manipulação de grandes conjuntos de dados na área de Bioinformática.

Qin e Fahringer [QIN07] também afirmam que aplicações científicas consomem uma porção de um conjunto de dados produzidos por qualquer aplicação e que um construtor para execuções em paralelo consome múltiplos conjuntos de dados em cada interação. Os autores afirmam, entretanto, que o problema que trata a especificação de como os conjuntos de dados e seus respectivos elementos podem ser identificados e como esses dados podem ser distribuídos dentro de interações em paralelo, ainda não foi totalmente explorado. Muitos sistemas de

workflow em grade resolvem o problema de replicação das entradas de conjuntos de dados para atividades, mas afirmam que existe a necessidade de mais estudos que busquem flexibilizar mecanismos de fluxo de dados de conjuntos de dados, evitando redundância. A abordagem definida por Qin e Fahringer [QIN07] reduz a duplicação de dados e otimiza a transferência entre atividades do *workflow*.

O trabalho desenvolvido nesta Tese não trabalha especificamente com processamento em grade, mas os conceitos aplicados no trabalho de Qin e Fahringer [QIN07] são utilizados na possibilidade do padrão ser executado em paralelo. Além disso, tanto o trabalho desenvolvido pelos autores, quanto o desenvolvido nesta Tese, manipulam grandes volumes de dados e o fazem a partir da definição de conjuntos.

8.3 Redes de Petri para Sistemas Biológicos

Chaouiya, em [CHA07], afirma que o uso de modelos matemáticos é crescente na representação de redes biológicas complexas. Destaca, principalmente, o uso das redes de Petri e suas extensões. Seu trabalho envolve a apresentação de como uma rede de Petri poderia representar um sistema biológico.

O autor apresenta um levantamento com diferentes trabalhos que utilizam redes de Petri para modelar aplicações na área científica. Seu trabalho enfatiza aspectos como efetividade na modelagem com redes de Petri e a análise e simulação das redes. O autor acredita que o aumento do uso de modelos baseados em redes de Petri para a representação de redes biológicas pode ser justificado pela representação gráfica do modelo, a possibilidade de representar sistemas concorrentes, sua base matemática e a existência de ferramentas de modelagem. Ainda afirma que, pela característica da aplicação da área Biológica envolver grandes interações de dados, existe a necessidade de se prover modelos qualitativos, os quais permitam uma análise formal.

O autor destaca como pontos principais na utilização de redes de Petri para a formalização de processos científicos:

- Redes de Petri possuem representação gráfica, com base teórica matemática e ferramentas de diagramação disponíveis;
- Redes de Petri permitem a análise de estruturas qualitativas para propriedades comportamentais quantitativas;
- Redes de Petri são efetivas para a modelagem de redes moleculares.

Apesar das redes de Petri terem sido definidas, inicialmente em 1962, sua utilização ainda é ampla, principalmente, quando se necessita de um formalismo com características matemáticas que represente o comportamento das tarefas (no caso de redes de Petri coloridas). A Tese aqui apresentada trabalha com redes de Petri coloridas, e o trabalho desenvolvido por Chaouiya [CHA07] embasa essa escolha. Apesar disso, diferente dos trabalhos analisados por Chaouiya, a modelagem em redes de Petri realizada nesta Tese é de um padrão, podendo ser utilizado por diferentes aplicações da área científica, desde que com características semelhantes.

8.4 Uma Linguagem de Fluxo de Dados baseada em Redes de Petri e Cálculo Relacional

Hidders et al., [HID08], propõem o que denominam de DFL: uma linguagem de *workflow* formal e gráfica para fluxos de dados. Justificam seu trabalho pela existência de *workflows* no qual grandes volumes de dados complexos são manipulados, e a estrutura desses dados reflete-se no *workflow*. Posicionam o trabalho desenvolvido como uma extensão das redes de Petri, sendo responsável pela organização do processamento das tarefas; e do cálculo relacional aninhado, com uma linguagem de consulta sobre objetos complexos, que também é responsável por manipular coleções de itens de dados.

Os autores afirmam que fluxos de dados são encontrados na prática, por exemplo, em experimentos *in silico* na Bioinformática e em sistemas que processem coleções de dados na Física, Astronomia e em outras ciências. A característica dessas áreas é possuir grandes volumes de estruturas de dados que são analisadas por um sistema e organizadas em rede, de forma que os fluxos de dados sejam processados. Para os autores, existem formalismos bem desenvolvidos para *workflows* que são baseados em redes de Petri. Entretanto, afirmam que esses formalismos não manipulam estruturas complexas de dados, de forma que reflitam na estrutura do *workflow* em questão. Em função disso, também utilizam o cálculo relacional. Neste trabalho, os autores apresentam a formalização da extensão realizada e, com o objetivo de validarem seu trabalho, mapearam um fluxo de dados real de Bioinformática, mesma área de aplicação da Tese aqui desenvolvida.

Este trabalho é relacionado ao desenvolvido nesta Tese pela área de aplicação e por trabalhar com redes de Petri. Além disso, por formalizar a extensão desenvolvida, da mesma forma que o feito nesta Tese para o P-MIA.

8.5 Abordagem para Concepção de Experimentos Científicos em Larga Escala Apoiados por *Workflows* Científicos

Pereira e Travassos, [PER09], afirmam que a ciência se apóia em infraestrutura computacional complexa para realizar pesquisas, interessando-se, principalmente, em estudos *in virtuo* e *in silico*, com a utilização de tecnologias de *workflow* científico. No seu trabalho, os autores propõem uma abordagem para auxiliar a concepção de *workflows* científicos para esses tipos de estudo.

Com o objetivo de capturar o conhecimento tácito envolvido na concepção de *workflows*, a abordagem direciona-se à identificação de requisitos e à modelagem do *workflow* nos níveis mais altos de abstração, independentemente de sistema gerenciador de *workflow* científico a ser utilizado. A abordagem apresentada pelos autores explora os conceitos do diagrama de atividades da UML [OMG10]. O trabalho de Pereira e Travassos, [PER09], apresenta relação com o trabalho apresentado nessa Tese, pela representação de *workflows* científicos e pela necessidade de apoiar o processamento computacional. Apesar disso, preocupa-se, apenas, com a modelagem de *workflows* científicos, automatizando o processo, sem se preocupar com a redução de dados a serem processados, foco principal do trabalho desenvolvido nesta Tese.

8.6 Padrões de Computação Paralela para *Workflows* em Grade

Pautasso e Alonso, em [PAU06], identificam um conjunto de padrões de *workflow* relacionados a execuções paralelas. Apresentam como os padrões podem ser representados em diferentes linguagens de *workflow* em grade e suas implicações para o projeto de execução de *workflows*. No trabalho desenvolvido pelos autores, o objetivo também é classificar padrões que manipulem paralelismo, direcionando-os à utilização em grade e não, necessariamente, identificar novos padrões. Para os autores, a execução em paralelo é uma técnica capaz de reduzir o tempo de execução de *workflows* científicos, pois conjuntos de tarefas que não tenham dependência podem ser executados em paralelo. Um conceito interessante, explorado pelos autores, é o do paralelismo adaptativo de dados, onde a quantidade de partições a serem executadas pode ser definida manualmente, ou em tempo de execução, e que a estrutura de uma instância de *workflow* não é somente influenciada pelos seus dados de entrada, mas pelas propriedades do ambiente de execução. A Tese desenvolvida está relacionada com o trabalho de Pautasso e Alonso [PAU06] por trabalhar com padrões de dados. Apesar disso, diferencia-se na característica do

padrão, de ser específico para manipulação de dados a serem submetidos ao processamento, e pela análise desses dados em tempo de execução determinar as próximas etapas.

8.7 Uma arquitetura de baixo acoplamento para execução de padrões de controle de fluxo em grades

Nardi, em [NAR09], justifica que o uso de padrões de *workflow* para controle de fluxo em aplicações de e-Science resulta em maior produtividade por parte do cientista, permitindo que se concentre em sua área de especialização. O autor também afirma que, apesar de todos os avanços, o uso de padrões de *workflow* para paralelização em grades permanece uma questão em aberto. Seu trabalho apresenta uma arquitetura de baixo acoplamento e extensível, que permite a execução de padrões com ou sem a presença de grade de modo transparente ao cientista, e a implementação de padrões de execução de *workflow*.

O autor define o Padrão Junção Combinada, já apresentado no documento desta Tese no Capítulo 4, que atende a diversos cenários de paralelização, comumente encontrados em aplicações de *e-Science*. Além disso, o autor definiu uma arquitetura, orientada a serviços, oferecendo flexibilidade e extensibilidade à solução.

A relação entre os dois trabalhos está na utilização de padrões amplamente difundidos, para a definição de outro padrão, e na busca por soluções que otimizem o processamento de dados científicos. A principal diferença está na utilização por parte de Nardi [NAR09] de padrões de fluxo e desta Tese de padrões de dados.

8.8 Redes de Petri como um Formalismo de Comparação

Grando et al. [GRA09] discutem a possibilidade de redes de Petri coloridas, como um formalismo, apoiam a análise de expressividade e verificação estrutural, comportamento e propriedades temporais em *workflows* clínicos (área médica). Os autores apresentam uma linguagem que pode ser formalmente mapeada para a representação de redes de Petri coloridas. Durante seu trabalho, os autores buscaram a formalização de algo informal, de orientações médicas baseadas em texto, por exemplo.

O relacionamento do trabalho desenvolvido pelos autores com o desta Tese está, apenas, na utilização de um formalismo matemático (redes de Petri coloridas), bem conhecido, com semânticas formais padronizadas, tendo uma representação gráfica padronizada e independente de fornecedor.

8.9 Métodos de Discretização e Discretização dos Dados de Docagem de Receptor Flexível

Machado et al. em [MAC10] e em [MAC10a] desenvolveram trabalhos também para o LABIO. Com o objetivo de reduzir demandas computacionais e descobrir informações sobre a interação entre receptores e ligantes, aplicaram diferentes algoritmos de mineração onde os arquivos de entrada são baseados na energia livre de ligação (FEB). Uma vez que o FEB apresenta um valor contínuo e os algoritmos de classificação necessitam de atributos categóricos, os autores também compararam três métodos de discretização para o FEB: por igualdade de frequência, por igualdade de largura e pela análise do desvio padrão. Além disso, avaliaram o impacto na geração de árvores de decisão.

Os autores identificaram que o método que apresentou melhor resultado após a aplicação dos algoritmos de mineração foi o que envolveu a análise do desvio padrão dos dados. Dessa forma, por mais que a técnica seja diferente e que a função utilizada não seja a mesma, os trabalhos relacionam-se pela utilização do desvio padrão, relacionado ao resultado médio obtido após o processamento dos *snapshots* e pela utilização do FEB como parâmetro de entrada.

8.10 Paralelismo de Dados em *Workflows* na Área de Bioinformática

Coutinho et al. em [COU10] definem que é muito comum, em experimentos de bioinformática, o processamento de grandes conjuntos de dados. Em função disso, afirmam que o paralelismo de dados é uma abordagem comum para incrementar o desempenho e reduzir o tempo de execução. Entretanto, afirmam que muitos sistemas gerenciadores de *workflows* científicos (SWfMS) suportam execuções paralelas apenas em ambientes computacionais de alto desempenho. Colocam o Hydra como um *middleware* com o propósito de ser uma ponte entre os SWfMS e os ambientes de alto desempenho, fornecendo um caminho transparente aos cientistas na paralelização de execuções de *workflows*. Seu trabalho analisa diferentes cenários de paralelismo de dados no domínio de Bioinformática e apresenta uma extensão específica para a manipulação de dados em paralelo em *workflows* na área.

O Hydra, conforme os autores, é um middleware que provê um conjunto de componentes capazes de serem incluídos em especificações de *workflow* de qualquer SWfMS para controlar o paralelismo de atividades. Dessa forma, esse trabalho relaciona-se com o definido nesta Tese, pela possibilidade de, após se implementar o P-MIA, criando seus componentes, poder ser configurada sua execução em paralelo com a utilização do Hydra, fornecendo um ganho ainda maior de processamento.

8.11 Um *Workflow* Científico para a Modelagem do Processo de Desenvolvimento de Fármacos Assistido por Computador Utilizando Receptor Flexível

Karina Machado em [MAC07a] e Machado et al. em[MAC07], apresentaram um *workflow* científico para subsidiar o processo de desenvolvimento de fármacos assistido por computador. Para a criação do *workflow* científico foram implementados *shell scripts* e programas que executassem efetivamente e de forma automática as sequências das atividades executadas pelo *workflow*. Esse modelo permitiu que docagens moleculares, que considerassem a flexibilidade explícita tanto do ligante como do receptor, fossem facilmente executadas, sendo então a flexibilidade natural das moléculas biológicas consideradas nos experimentos de docagem. Além disso, com a automatização do processo, foi possível a inclusão de uma etapa de seleção de *snapshots* para que não fosse necessária a execução de todas as conformações do receptor, reduzindo o tempo necessário para se analisar a interação receptor-ligante. Essa etapa, entretanto, somente pode ser utilizada após a realização de um experimento exaustivo, selecionando aqueles *snapshots* da macromolécula que apresentaram os melhores resultados de docagem com determinado ligante. Esses *snapshots* são utilizados para a docagem de outros ligantes pertencentes à mesma classe que o primeiro. Outros detalhes do trabalho desenvolvido por Karina Machado também foram apresentados dos Capítulos 3 e 6.

O trabalho desenvolvido nesta Tese relaciona-se com o desenvolvido por Machado et al. por trabalhar com *workflows* científicos, na área de Bioinformática, e pelos dados da aplicação teste serem provenientes de experimentos do LABIO. Além disso, conforme já apresentado no Capítulo 6, o trabalho desenvolvido nesta Tese substitui etapas do *workflow* desenvolvido por Karina, buscando a redução da quantidade total de experimentos a serem executados e, se possível, reduzindo o tempo total de processamento. Apesar disso, com o padrão definido nesta Tese, busca-se a aplicação em diferentes áreas científicas, desde que com características semelhantes à Bioinformática.

8.12 Considerações do Capítulo

Este capítulo apresentou trabalhos relacionados ao desenvolvido nesta Tese e, buscando sintetizar esses trabalhos, comparando-os com o desenvolvido na Tese, a Tabela 32 foi elaborada. Analisando-se a Tabela 32, verifica-se o relacionamento entre os trabalhos pesquisados e o P-MIA, identificando suas principais características. Os critérios de comparação apresentados nessa tabela, além de alguns serem conceitos amplamente difundidos na literatura, também envolvem características de áreas científicas. São eles:

- *Padrões de Dados*: são identificados os trabalhos relacionados que utilizam conceitos de padrões de dados, que definem novos padrões ou que reconhecem sua importância, uma vez que o trabalho desenvolvido nesta Tese define um novo padrão de dados.
- *Padrões de Fluxo*: são identificados os trabalhos relacionados que utilizam conceitos de padrões de fluxos, que definem novos padrões ou que reconhecem sua importância, uma vez que o trabalho desenvolvido nesta Tese, apesar de não manipular especificamente padrões de fluxo, os utiliza como base para sua definição, como é o caso do padrão de fluxo desenvolvido por Nardi [NAR09].
- *Paralelismo*: são identificados os trabalhos que estudam conceitos de paralelismo aplicados a dados, independente da forma de aplicação, bem como os que reconhecem sua utilização. Por mais que o padrão de dados desenvolvido nesta Tese não utilize explicitamente conceitos de paralelismo, o padrão pode ser implementado sobre esse conceito.
- *Execução em Grade*: são identificados os trabalhos que estudam conceitos específicos de execução em grade, sendo essa uma das possibilidades de manipulação de dados em paralelo. O trabalho de Nardi [NAR09] é definido sobre esse conceito e, uma vez que se afirma que o padrão definido nesta Tese possui como base o trabalho por ele desenvolvido, esses conceitos relacionam-se.
- *Formalização em redes de Petri coloridas*: todos os trabalham que mencionam a utilização de redes de Petri ou de redes de Petri coloridas aplicados à área científica são destacados, uma vez que seus conceitos são utilizados para a formalização do padrão definido nesta Tese.
- *Workflows científicos*: o padrão definido nesta Tese é desenvolvido especificamente para utilização em *workflows* científicos, pela característica do processo.
- *Grandes volumes de dados*: uma das características dos *workflows* científicos é a manipulação de grandes volumes de dados e essa é, também, a principal característica das atividades desenvolvidas pelo LABIO, quando se busca redução da quantidade de processamento e otimização dos processos.

- *Adaptação*: os conceitos de adaptação são importantes por uma das características do padrão: definir, dinamicamente, quais os próximos *snapshots* a serem executados. Dessa forma, esse critério relaciona-se com o padrão definido.

Nota-se, ao analisar a Tabela 32, que os diversos trabalhos contribuem para que o P-MIA agregue as características dos critérios definidos, uma vez que serviram como base para seu desenvolvimento. Dessa forma, atende-se à necessidade de uma área que trabalhe com padrões de dados, manipule grandes volumes de informações, crie e gerencie *workflows* científicos, possibilite processamento em paralelo, ou em grade, e trabalhe com conceitos de adaptação em tempo de execução.

Tabela 32 – Síntese e comparação entre os trabalhos relacionados e o P-MIA

	Padrões de Dados	**Padrões de Fluxos**	**Paralelismo**	**Execução em Grade**	**Formalização em redes de Petri coloridas**	***Workflows* Científicos**	**Grandes volumes de dados**	**Adaptação**
Fluxos de Dados e Validações em Modelagens de *Workflows* [SAD04]	Não trabalham explicitamente com padrões de Dados, mas com fluxos de dados	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Sim, quando direciona seus estudos à Bioinformática	Sim, quando direciona seus estudos à Bioinformática	Não se aplica
Suporte a Fluxos de Dados Avançados para Aplicações de *Workflows* Científicos em Grade [QIN07]	Não trabalham explicitamente com padrões de Dados, mas com fluxos de dados	Não se aplica	Execução em Grade	Trabalham com a execução de *workflows* científicos em grade	Não se aplica	Trabalham com o conceito de *workflows* científicos	Trabalham com a possibilidade de manipular grandes volumes de dados	Não se aplica
Redes de Petri para Sistemas Biológicos [CHA07]	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Utiliza rede de Petri para representar um sistema biológico	Não se aplica	Não se aplica	Não se aplica
Uma Linguagem de Fluxo de Dados baseada em Redes de Petri e Cálculo Relacional [HID08]	Não trabalham explicitamente com padrões de Dados, mas com fluxos de dados	Não se aplica	Não se aplica	Não se aplica	Apresentam uma extensão das redes de Petri	Não se aplica	Trabalham com a possibilidade de manipular grandes volumes de dados	Não se aplica
Abordagem para Concepção de Experimentos Científicos em Larga Escala Suportados por *Workflows* Científicos [PER09]	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Representação de *workflows* científicos, buscando apoiar o processamento computacional	Não se aplica	Não se aplica
Padrões de Computação Paralela para *Workflows* em Grade [PAU06]	Classificação de padrões que manipulam paralelismo	Não se aplica	Trabalham com paralelismo de dados	Trabalham com o conceito de *workflow* em grade	Não se aplica	Não se aplica	Não se aplica	Exploram o conceito de paralelismo adaptativo de dados
Uma arquitetura de baixo acoplamento para execução de padrões de controle de fluxo em grades [NAR09]	Não se aplica	Usa padrões de *workflow* para controle de fluxo em aplicações de e-Science	O padrão desenvolvido atende a diferentes cenários de paralelização	Permite a execução de padrões com ou sem a presença de grade	Utiliza redes de Petri para representar graficamente seu padrão	Utiliza-se do conceito de *workflows* científicos	Não menciona claramente, mas seus esforços direcionam a solução à manipulação de grandes volumes de dados	Não se aplica
Redes de Petri como um Formalismo de Comparação	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Utiliza redes de Petri coloridas	Trabalha com o conceito de	Não se aplica	Não se aplica

[GRA09]					como um formalismo	*workflows* científicos		
Métodos de Discretização e Discretização dos Dados de Docagem de Receptor Flexível [MAC10], [MAC10a]	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Trabalham com mineração sobre grandes volumes de dados	Não se aplica
Paralelismo de Dados em *Workflows* na Área de Bioinformática [COU10]	Não se aplica	Não se aplica	Trabalham com paralelismo de dados	Não explicitamente, mas se aplica	Não se aplica	Desenvolveram um middleware com o propósito de ser uma ponte entre os SWfMS e os ambientes de alto desempenho	Trabalham com grandes volumes de dados	Não se aplica
Um *Workflow* Científico para a Modelagem do Processo de Desenvolvimento de Fármacos Assistido por Computador Utilizando Receptor Flexível [MAC07], [MAC07], [MAC07a]	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Não se aplica	Implementou um *workflow* científico, buscando automatizar processos antes manuais	Processa grandes volumes de dados, apresentando, apesar da automatização, ainda grande tempo de processamento	Não se aplica
P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis: Um Padrão de Dados para *Workflows* Científicos	Define um padrão de dados para a autoadaptação de instâncias em tempo de execução	Utiliza como modelo o padrão de controle de Fluxo desenvolvido por Nardi em [NAR09]	O padrão definido pode ser executado em paralelo, com diferentes conjuntos de dados	O padrão definido pode ser executado em estruturas em grade, desde que controlado por ambiente específico	O padrão é representado por meio de redes de Petri coloridas	O padrão pode ser inserido em qualquer implementação de *workflow* científico, desde que atenda às características da área.	O padrão atende à necessidade de manipulação de grandes volumes de dados, trabalhando com o conceito de conjuntos	O padrão apresenta conceitos de autoadaptação, definindo os próximos dados a serem instanciados em tempo de execução

9 CONSIDERAÇÕES FINAIS

A área que estuda *workflows* não é uma área recente. Estudos têm sido realizados desde a década de 80 e, na década de 90, surgiu a necessidade de uma maior padronização e definição de tipos desses processos [GEO95, HOL95, LEY00]. Este documento apresentou estudos sobre *workflows* de negócios, *workflows* científicos, padrões que direcionam a implementação de processos nesses sistemas, formalismos utilizados para a descrição dos processos, bem como a área de aplicação, que propiciou a realização dos experimentos, executados após a definição e formalização do padrão de dados P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis.

O Capítulo 2 apresentou o estudo realizado sobre *workflows* de negócios e *workflows* científicos, identificando diferenças entre as duas abordagens, bem como formalismos utilizados para a representação de processos: redes de Petri e redes de Petri coloridas. Conforme [AAL02], a utilização de um conceito formal propicia algumas vantagens e faz com que a definição de um processo seja mais precisa. A formalização pode diminuir ambiguidades, incertezas e contradições. Além disso, facilita a representação de sistemas complexos, bem como do comportamento desses sistemas [FOK00].

No Capítulo 3 a área de Bioinformática foi apresentada e as atividades realizadas pelo LABIO. Alguns padrões de dados definidos na literatura foram apresentados no Capítulo 4 e foram utilizados como base para a definição do padrão Múltiplas Instâncias Autoadaptáveis – P-MIA, definido nesta Tese de Doutorado, visando atender à necessidade de autoadaptação de instâncias em execução, com a manipulação de um grande volume de dados. Conceitos de adaptação e autoadaptação foram discutidos por Hübler e Ruiz em [HUB07] e em [HUB09] e, por serem apenas aplicados na instanciação dos dados no padrão, não foram detalhados nesta Tese.

No Capítulo 5 a formalização do P-MIA foi apresentada, contendo as principais definições, bem como sua representação gráfica, realizada com a ferramenta CPN Tools. Seu funcionamento foi detalhado no Capítulo 6, enfatizando as regras e características do padrão. Os testes com dados reais foram detalhados no Capítulo 7 e os trabalhos relacionados foram confrontados com o trabalho desenvolvido nesta Tese no Capítulo 8.

9.1 Principais Contribuições

Ao se considerar que uma das principais características de áreas como a Bioinformática é a manipulação de grandes volumes de dados e o processamento desses dados no menor espaço de tempo possível, o P-MIA, apresentado neste documento, atende à questão de pesquisa previamente definida: "Como reduzir a quantidade de *snapshots* a serem processados, reduzindo o tempo total de processamento e procurando manter o mesmo nível de acerto na identificação de compostos promissores?".

O padrão definido atende à questão de pesquisa, ao propiciar que se alcance o objetivo geral definido nesta Tese: “melhorar o tempo final de processamento, reduzindo a quantidade de experimentos de docagem molecular, com base nos resultados obtidos em tempo de execução”. O padrão manipula conjuntos de dados, previamente agrupados por similaridade e, com base em resultados obtidos por elementos desses conjuntos, define a execução ou não dos demais elementos. Dessa forma, nem todos os experimentos serão realizados, reduzindo o tempo final de processamento e mantendo um patamar de qualidade.

Quanto aos objetivos definidos como específicos:

- *Definir um padrão de dados capaz de ser utilizado pela área de Bioinformática e por outras áreas que apresentem características semelhantes*: o padrão definido P-MIA, formalizado no Capítulo 5 e detalhado no Capítulo 6, apesar de utilizar dados específicos de Bioinformática para a realização dos experimentos, apresentou características que podem ser facilmente aplicadas em outras áreas científicas. Para essa comprovação, entretanto, seria interessante que experimentos com dados de outras áreas fossem realizados.

- *Definir uma função que não descarte dados que apresentem a probabilidade de serem promissores*: esse objetivo foi alcançado ao se adaptar a função da Regra Empírica, denominada neste trabalho de Regra Empírica Adaptada, apresentada no Capítulo 6, servindo como subsídio para o cálculo da média estimada. Com o valor gerado a partir da média estimada, os *snapshots* não são descartados sem que apresentem a probabilidade de obterem bons resultados, pois a média estimada leva em consideração o desvio padrão dos resultados.

- *Reduzir a quantidade total de dados a serem processados*: o princípio fundamental do P-MIA é a redução da quantidade de *snapshots* a serem processados. Os testes realizados no Capítulo 7 demonstram que se obtém um ganho, com a utilização do

padrão, podendo chegar, quando a análise inicia aos 20% do processamento dos *snapshots* concluído, com o ligante PIF a 47% de ganho e com o ligante NADH a 43% de ganho.

- *Buscar manter a qualidade dos dados processados*: com os testes realizados no Capítulo 7 verificou-se que a grande maioria dos *snapshots* que apresentaram melhores resultados de processamento foi processada, também, por meio do P-MIA. Entretanto, acredita-se que esse objetivo seria alcançado com um índice ainda maior se o critério de similaridade, para o agrupamento dos dados, fosse específico para a realidade implementada.

Neste contexto, com a definição de um padrão de dados capaz de inferir maior velocidade de processamento aos experimentos, com a formalização do padrão e a descrição em detalhes do seu funcionamento e dos testes realizados, acredita-se que o P-MIA é uma contribuição interessante para a comunidade científica.

9.2 Trabalhos Futuros

Como sugestões de trabalhos futuros:

- Implementação do P-MIA: Padrão Múltiplas Instâncias Autoadaptáveis, na forma de um componente, possibilitando sua utilização por diferentes aplicações de *workflow* científico. Essa implementação é assunto em desenvolvimento na Dissertação de Mestrado de Fábio Frantz, em parceria com o LABIO.
- Estudos de viabilidade de execução do padrão em diferentes ambientes de alto desempenho, junto ao Laboratório de Alto Desempenho da PUCRS: já em desenvolvimento pela Mestranda Renata de Paris, também em parceria com o LABIO.
- Definição de uma função de similaridade que seja específica para a realidade estudada também tende a fornecer melhores resultados. Essa função de similaridade já está em definição, sendo foco do trabalho de Doutorado de Karina Machado. Assim, após a definição dessa função, novos testes devem ser realizados, buscando evidenciar o ganho ou não de processamento.
- Aplicação do padrão com dados de outras áreas científicas, diferentes da Bioinformática, mas com características semelhantes.

REFERÊNCIAS

[AAL01] AALST, W.M.P.V.D; KUMAR, A. "A reference model for team-enabled *workflow* management systems". *Data & Knowledge Engineering*, vol. 38-3, September 2001, pp. 335-363.

[AAL02] AALST, W.M.P.V.D.; HEE, K. V. "*Workflow* Management: Models, Methods, and Systems". Cambridge, Massachusetts: The MIT Press, 2002, 368p.

[AAL03] AALST, W.M.P.V.D.; HOFSTEDE, A.H.; WESKE, M. "Business Process Management: A Survey". In: 1st International Conference on Business Process Management, 2003, pp. 1-12.

[AAL03a] AALST, W.M.P.V.D.; HOFSTEDE, A.H.M.; KIEPUSZEWSKI, B.; BARROS, A.P. "*Workflow* Patterns". *Distributed and Parallel Databases*, vol. 14-1, 2003, pp. 5-51.

[AAL07] AALST, W.M.P.V.D. "*Workflow* Patterns". Capturado em: http:// www.workflowpatterns.com, Junho 2007.

[AAL98] AALST, W.M.P.V.D. "The Application of Petri Nets to *Workflow* Management". *The Journal of Circuits, Systems and Computers*, vol. 8-1, 1998, pp. 1-53.

[BAN94] BANERJEE, A.; DUBNAU, E.; QUEMARD, A.; BALASUBRAMANIAN, V.; UM, K.; WILSON, T.; COLLINS, D.; DE LISLE, G.; JACOBS JR., W. "InhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis". *Science* 263, 1994, pp. 227–230.

[BER00] BERMAN, H. M.; WESTBROOK, J. ; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N. ; BOURNE, P. E. "The Protein Data Bank". *Nucleic Acids Research*. vol. 28-1, 2000, pp. 235-242.

[BRA06] BRAGHETTO, K. R. "Padrões de Fluxo de Processos em Banco de Dados Relacionais." Dissertação de Mestrado em Ciência da Computação, Programa de Pós-Graduação em Ciência da Computação, IME-USP, 2006, 126p.

[CHA07] CHAOUIYA, C. "Petri net modelling of biological networks". *Brief Bioinform*, vol. 8-4, 2007, pp. 210 – 219.

[COU10] COUTINHO, F.; OGASAWARA, E.; DE OLIVEIRA, D.; BRAGANHOLO, V.; LIMA, A. A.; DÁVILA, A. M.; MATTOSO, M. "Data parallelism in bioinformatics *workflows* using Hydra". In: 19th ACM international Symposium on High Performance Distributed Computing, 2010, pp. 507-515.

[CPN10] CPN Group, University of Aarhus, Denmark. "CPN Tools Home Page". Capturado em http://wiki.daimi.au.dk/cpntools/, Outubro 2010.

[DAV08] DAVIDSON, S.; FREIRE, J. "Provenance and scientific *workflows*: Challenges and opportunities". In: SIGMOD Conference, 2008, pp. 1345-1350.

[DEE06] DEELMAN, E.; GIL, Y. “Managing large-scale scientific *workflows* in distributed environments: Experiences and challenges”. In: E-SCIENCE '06 - Proceedings of the Second IEEE International Conference on e-Science and Grid Computing, 2006, 6p.

[DEE09] DEELMAN, E.; GANNON, D.; SHIELDS, M.; TAYLOR, I. “*Workflows* and e-science: An overview of *workflow* system features and capabilities”. *Future Generation Computer Systems*, vol. 25-5, 2009, pp. 528-540.

[DES95] DESSEN, A.; QUÉMARD, A.; BLANCHARD, J.S.; JACOBS, W.R.; JR SACCHETTINI, J.C. “Crystal structure and function of the isoniazid target of Mycobacterium tuberculosis”. *Science,* vol. 267-5204, 1995, pp. 1638-1641.

[DEV06] DEVORE, J. L. “Probabilidade e Estatística para Engenharia e Ciências”. São Paulo: Pioneira Thomson Learning, 2006, 6ª Edição, 692p.

[EWI01] EWING, T. J. A.; MAKINO, S.; SKILLMAN, A. G.; KUNTZ, I. D. “DOCK4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases”. *Journal of Computer-Aided Molecular Design*, vol. 15-5, 2001, pp. 411–428.

[FOK00] FOKKING, W. J. “Introduction to Process Álgebra: Texts in Theoretical Computer Science”. Berlin: Springer – Verlag, 2000. 163p.

[GEO95] GEORGAKOPOULOS, D.; HORNICK, M. F.; SHETH, A. P. “An overview of *workflow* management: From process modeling to *workflow* automation infrastructure”. *Distributed and Parallel Databases,* vol. 3-2, 1995, pp. 119-153.

[GIL07] GIL, Y.; DEELMAN, E.; ELLISMAN, M.; FAHRINGER, T.; FOX, G.; GANNON, G.; GOBLE, C.; LIVNY, M.; MOREAU, L.; MYERS, J. “Examining the Challenges of Scientific *Workflows*”. *IEEE Computer*, vol. 40-12, 2007, pp. 24-32.

[GIR02] GIRAULT, C.; VALK, R. “Petri Nets for Systems Engineering: A Guide for Modeling, Verification and Application”. Springer Verlag, 2002, 607p.

[GLA08] GLATARD, T.; MONTAFNAT, J.; LINGRAND, D.; PENNEC, X. “Flexible and efficient *workflow* deployment of data-intensive applications on grids with MOTEUR”. *International Journal of High Performance Computing and Applications,* vol. 22-3, 2008, pp. 347-360.

[GOO96] GOODSELL, D.; MORRIS, G.; OLSON, A. “Docking of Flexible Ligands: Applications of AutoDock”. *Journal of Molecular Recognition*, vol. 9-1, 1996, pp. 1–5.

[GRA09] GRANDO, M.A.; GLASSPOOL, D.W.; FOX, J. “Petri nets as a formalism for comparing expressiveness of *workflow*-based clinical guideline languages”. *Lecture Notes in Business Information Processing*, vol. 17-5, 2009, pp. 348-360.

[GUB06] GUBALA, T.; HAREZLAK, D.; BUBAK, M.; MALAWSKI, M. “Semantic composition of scientific *workflows* based on the petri nets formalism”. In: 2nd IEEE International Conference on e-Science and Grid Computing, 2006, 8p.

[HEU88] HEUSER, C. A. “Modelagem Conceitual de Sistemas”. Buenos Aires, Argentina: Editorial Kapeluz S.A., 1988, 94p.

[HID08] HIDDERS, J.; KWASNIKOWSKA, N.; SROKA, J.; TYSZKIEWICZ, J.; BUSSCHE, J. “Dfl: A dataflow language based on Petri nets and nested relational calculus”. *Information Systems*, vol. 33-3, 2008, pp. 261–284.

[HOL95] HOLLINSWORTH, D. “The *Workflow* Reference Model”. *Workflow* Management Coalition, Technical Report TC00-1003,1995, 55p.

[HOS00] HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WOHLIN, C.; RUNESON, P.; WESSLÉN, A. “Experimentation in software engineering: an introduction”. Norwell: Kluwer Academic Publishers, 2000, 1st edition, 204p.

[HUB07] HUBLER, P. N.; RUIZ, D. D. “Auto-Adaptabilidade de *Workflows* representada através da Álgebra de Processos.” In: XXXIII Conferencia Latinoamericana De Informatica, 2007, 10p.

[HUB09] HUBLER, P. N.; RUIZ, D. D. “Mapping of Changes in Basic Control-Flow Patterns Using Process Algebra.” In: III Workshop de Gestão de Processos de Negócio (WBPM) do WebMedia 2009, 6p.

[IRW05] IRWIN, J. J.; SHOICHET, B. K. “ZINC – A Free Database of Commercially Available Compounds for Virtual Screening”. *Journal of Chemical Information and Modeling*, vol. 45-1, 2005, pp. 177-182.

[JAB96] JABLONSKI, S.; BUSSLER, C. “*Workflow* Management: Modeling Concepts, Architecture and Implementation.” International Thomson Computer Press, September 1996. 351p.

[JAI99] JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. “Data clustering: A review”. *ACM Computing Surveys*, vol. 31-3, 1999, pp. 264-323.

[JEN94] JENSEN, K. “An Introduction to the Theoretical Aspects of Coloured Petri Nets - A Decade of Concurrency”. *Lecture Notes in Computer Science*, vol. 803, 1994, pp. 230-272.

[JEN97] JENSEN, K. “A brief introduction to coloured Petri nets”. In: Tools and Algorithms for the Construction and Analysis of Systems - TACAS’97 Workshop, 1997, pp. 203–208.

[JUN10] JUNIOR, N. N. T. “Modelo–E10: Um Modelo para Estimativas de Esforço em Manutenção de Software”. Tese de Doutorado em Ciência da Computação, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2010, 134p.

[KOT05] KOTB, Y. T.; BAUMGART, A. S. “An Extendted Petri net for modeling *workflow* with critical sections”. In: IEEE International Conference on e-Business Engineering, 2005, 8p.

[KUN92] KUNYZ, I. D. “Structure-based Strategies for Drug Design and Discovery.” *Science*, vol. 257, 1992, pp. 1078–1082.

[KUO03] KUO, M. R.; MORBIDONI, H. R.; ALLAND, D.; SNEDDON, S. F.; GOURLIE, B. B.; STAVESKI, M. M.; LEONARD, M.; GREGORY, J. S.; JANJIGIAN, A. D.; YEE, C.; KREISWIRTH, B.; IWAMOTO, H.; PEROZZO, R.; JACOBS, W. R. Jr.; SACCHETTINI, J. C.; FIDOCK, D. A. “Targeting tuberculosis and malaria through inhibition of Enoyl reductase: compound activity and structural data”. *J. Biol. Chem*., vol. 278, 2003, pp. 20851–20859.

[LAR09] LARSON, R; FARBER B. “Estatística Aplicada”. São Paulo: Editora Pearson, 4º. Edição, 2009, 476p.

[LEY00] LEYMANN, F.; ROLLER, D. “Production *workflow*: concepts and techniques.” Prentice Hall, 2000. 479 p.

[LIN02] LIN, J. H.; PERRYMAN, A. L.; SCHAMES, J. R.; McCAMMON, J. A. “Computational drug design accommodating receptor flexibility: the relaxed complex scheme”. *J. Am. Chem. Soc*, vol. 124, 2002, pp. 5632-5633.

[LUD06] LUDÄSCHER, B.; ALTINTAS, I.; BERKLEY, C.; HIGGINS, D. ; JAEGER-FRANK, E.; JONES, M.; LEE, E.; TAO, J.; ZHAO, Y., "Scientific *Workflow* Management and the Kepler System". *Concurrency and Computation: Practice & Experience*, vol.18-10, 2006, pp. 1039-1065.

[LUD09] LUDÄSCHER, B.; WESKE, M.; MCPHILLIPS, T.; BOWERS, S. "Scientific *Workflows*: Business as Usual?". *Lecture Notes in Computer Science*, vol. 5701, 2009, pp. 31-47

[LUS01] LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. "What is Bioinformatics?: a proposed definition and overview of the field". *Methods of Information in Medicine*, vol.4, 2001, pp. 346-358.

[MAC07] MACHADO, K. S.; SCHROEDER, E. K.; RUIZ, D D.; SOUZA, O.N. "Automating molecular docking with explicit receptor flexibility using scientific *workflows*". In: 2nd Brazilian Symposium On Bioinformatics (Lecture Notes in Computer Science (4643)), 2007, p. 1-11.

[MAC07a] MACHADO, K. S. "Um *Workflow* Científico para a Modelagem do Processo de Desenvolvimento de Fármacos Assistido por Computador Utilizando Receptor Flexível." Dissertação de Mestrado em Ciência da Computação, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2007, 75p.

[MAC10] MACHADO, K. S. ; WINCK, A. T. ; RUIZ, D. D. ; NORBERTO DE SOUZA, O. "Discretization of Flexible-Receptor Docking Data". In: Brazilian Symposium on Bioinformatics (Advances in Bioinformatics and Computational Biology - LNBI-LNCS, vol. 6268), 2010, pp. 75-79.

[MAC10a] MACHADO, K. S. ; WINCK, A. T. ; RUIZ, D. D. ; NORBERTO DE SOUZA, O. "Comparison of Discretization Methods of Flexible-Receptor Docking Data for Analyses by Decision Trees". In: IADIS International Conference Applied Computing, 2010, pp.75-79.

[MAC96] MACIEL, P.; LINS, R.; CUNHA, P. "Introdução às redes de Petri e aplicações". Campinas: Instituto de computação - UNICAMP, 1996.

[MAT08] MATTOSO, A.; MATTOS, A.; SILVA, F., C.; RUBERG, N; CRUZ, S.M.S. "Gerência de *Workflows* Científicos: uma análise crítica no contexto da bioinformática". COPPE/UFRJ, PESC, Technical Report ES-716/08, 2008, 87p.

[MAT09] MATTOSO, M.; WERNER, C.; TRAVASSOS, G.; BRAGANHOLO, V.; MURTA, L.; OGASAWARA, E.; OLIVEIRA, F.; MARTINHO, W. "Desafios no Apoio à Composição de Experimentos Científicos em Larga Escala". In: *SEMISH - CSBC*, 2009, 15p.

[MCG06] MCGOUGH, A. S.; COHEN, J.; DARLINGTON, J.; KATSIRI, E.; LEE, W.; PANAGIOTIDI, S.; PATEL, Y. "An end-to-end workflow pipeline for large-scale grid computing". *Journal Grid Computing*, vol. 3-4, 2005, pp.259.281.

[MEY04] MEYER, L.; RÖSSLE, S.; BISCH, P.; MATTOSO, M. "Parallelism in Bioinformatics *Workflows*". In: Proceedings of the International Conference VECPAR, 2004, pp. 705–718.

[MUR89] MURATA, T. "Petri net: properties, analysis and applications". *Proceedings of the IEEE*, vol. 77-4, 1989, pp. 541-579.

[NAR09] NARDI, A. R. "Uma arquitetura de baixo acoplamento para execução de padrões de controle de fluxo em grades". Tese de Doutorado em Ciência da Computação, Programa de Pós-Graduação em Ciência da Computação, IME-USP, 2009, 162p.

[OLI04] OLIVEIRA, J.; SOUZA, E.; BASSO, L.; PALACI, M.; DIETZE, R.; SANTOS, D.; MOREIRA, I. "An inorganic iron complex that inhibits wild-type and an isoniazid-resistant mutant 2-transenoyl-ACP (CoA) reductase from Mycobacterium tuberculosis". *Chem. Commun*., vol. 3, 2004, pp. 312–313.

[OMG09] OBJECT MANAGEMENT GROUP. "Business Process Modeling Notation (BPMN)". Version 1.2. Technical report. Capturado em http://www.omg.org/spec/BPMN/1.2, Fevereiro 2009.

[OMG10] OBJECT MANAGEMENT GROUP. "OMG Unified Modeling Language Specification", versão 2.2, formal/09-02-02. Capturado em: http://www.omg.org/spec/UML/2.2/, Outubro 2010.

[PAU06] PAUTASSO, C.; ALONSO, G. "Parallel Computing Patterns for Grid *Workflows*." In: Proceedings of the Workshop on *Workflows* in Support of Large-Scale Science, 2006, pp. 1-10.

[PEA95] PEARLMAN, D. A.; CASE, D. A.; CALDWELL, J. W.; ROSS, W. R.; CHEATMAN, T. E.; DeBOLT, S.; FERGUSON, D.; SEIBEL, G.; KOLLMAN, P. AMBER, "A computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidade the structures and energies of molecules." *Computer and Physics Communication*, vol.91, 1995, pp. 1-41.

[PEN04] PENHA, D. O.; FREITAS, H.; MARTINS, C. A. P. S. "Modelagem de Sistemas Computacionais usando Redes de Petri: aplicação em projeto, análise e avaliação." Capturado em: http://www.inf.pucminas.br/professores/cota/papers/erirjes2004.pdf, Setembro 2009.

[PER09] PEREIRA, W. M.; TRAVASSOS, G. H. "Abordagem para concepção de experimentos científicos em larga escala suportados por *workflows* científicos". In: 3 E-Science Workshop colocado ao SBBD/SBES, 2009, pp. 25-32.

[PET62] PETRI, C. A. "Kommunikation mit Automaten." Ph.D. dissertation, Institut für instrumentelle Mathematik, Bonn, 1962.

[PLE02] PLESUMS, C. "Introduction to *Workflow*". *Workflow Handbook 2002*, 2002, pp. 19-38.

[PLE03] PLESUMS, C. "Getting Started with *Workflow*". *Workflow Handbook 2003*, 2003, pp. 257-262.

[PRI03] PRIOR, C. "*Workflow* and Process Management". *Workflow Handbook 2003*, 2003, pp. 17-25.

[QIN07] QIN, J.; FAHRINGER T. "Advanced data flow support for scientific grid *workflow* applications". In: Proceedings of the ACM/IEEE conference on Supercomputing (SC), 2007, pp. 1–12.

[RAR96] RAREY, M.; KRAMER, B.; LENGAUER, T.; KLEBE, G. "A Fast Flexible Docking Method Using an Incremental Construction Algorithm". *J. Mol. Biol*., vol.261, 1996, pp. 470–489.

[ROD08] RODRIGUEZ, M. C.; PRIOL, T.; NEMETH, Z. "Dynamicity in Scientific *Workflows*." CoreGRID Technical Report Number TR-0162. Capturado em: http://www.coregrid.net/mambo/images/stories/TechnicalReports/tr-0162.pdf Agosto 2008.

[RUS04] RUSSELL, N.; TER HOFSTEDE, A.; EDMOND, D.; AALST, W. "*Workflow* data patterns." Technical Report FIT-TR-2004-01, Queensland Univ. of Techn., 2004, 75p.

[RUS05] RUSSELL, N.; TER HOFSTEDE, A.; EDMOND, D.; AALST, W. "*Workflow* Data Patterns: Identification, Representation and Tool Support." In: Proc. of the 24th International Conference on Conceptual Modeling (ER 2005), 2005, pp. 353–368.

[RUS06] RUSSELL, N.; TER HOFSTEDE, A.; AALST, W; MULYAR, N. "*Workflow* Control-Flow Patterns - A Revised View". s.l. : BPM Center Report, BPM-06-22, 2006, pp. 6-22.

[SAD04] SADIQ S.W.; ORLOWSKA M.E.; SADIQ W.; FOULGER C. "Data Flow and Validation in *Workflow* Modelling". In: Fifteenth Australasian Database Conference (ADC), 2004, pp. 207-214.

[SAN02] SANT'ANNA, C.M.R. "Glossário de Termos Usados no Planejamento de Fármacos". *Quim. Nova*, vol.25-3, pp. 505-512.

[SCH05] SCHROEDER, E. K; BASSO, L. A.; SANTOS, D. S.; SOUZA, O. N. "Molecular dynamics simulation studies of the Wild-Type, I21V, and I16T Mutants of Isoniazid-Resistant Mycobacterium tuberculosis Enoyl Reductase (InhA) in Complex with NADH - Toward the Understanding of NADH-InhA Different Affinities". *Biophys.* J, vol.89, 2005, pp. 876-884,

[SCH91] SCHAEL, T. et al., "Design Principles for Cooperative Office Support Systems in Distributed Process Management", in Support Functionality in the Office Environment, A. Verrijn-Stuart (ed), North Holland, 1991. Capturado em: http://www.inf.ufrgs.br/~palazzo/docs/read/*workflow*.htm#caracterização

[SIN06] SINGH, J.; DENG, Z.; NARALE, G.; CHUAQUI, C. "Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein–small molecule complexes." *Chem Biol Drug Des*., vol.67, 2006, pp. 5–12.

[TAY06] TAYLOR, I.J.; DEELMAN, E.; GANNON, D.B.; SHIELDS, M. (Eds.) "*Workflows* for e-Science: Scientific *Workflows* for Grids", Springer Verlag, 2006, 530p.

[TRA03] TRAVASSOS, G. H.; BARROS, M. O. "Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering". In: Workshop on Empirical Software Engineering - The Future of Empirical Studies in Software Engineering, 2003, pp. 117-130.

[WES95] WESKE, M.; VOSSEN, G.; MEDEIROS, C. "Scientific *Workflow* Management: WASA Architecture and Applications". In: Proceedings of 6th DEXA Conference, 1995, pp. 574–583.

[WIN09] WINCK, A. T. ; MACHADO, K. ; SOUZA, O. N. ; RUIZ, D. D. "FReDD: supporting mining strategies though a flexible-receptor docking database". In: IV Brasilian Symposium on Bioinformatics, 2009, pp. 143-146.

[WIN10] WINCK, A. T. ; MACHADO, K. S. ; NORBERTO DE SOUZA, O. ; RUIZ, D. D. "Supporting Intermolecular Interaction Analyses of Flexible-Receptor Docking Simulations". In: IADIS International Conference Applied Computing, 2010, pp. 183-190.

[WOR06a] WORLD Health Organization. "Global tuberculosis control: surveillance, planning, financing." WHO Report . 2006. 255p.

[WOR06b] WORLD Health Organization. "Tuberculosis facts." Capturado em: http://www.who.int/tb/publications/2006/tb_facts_2006.pdf, Outubro 2010.

[WOR99] WORKFLOW Management Coalition. “The *Workflow* Management Coalition Specification – Terminology & Glossary”. Document Number WFMC-TC-1011, Feb 1999, 65p.

[ZIN09] ZINC. “The University of California at San Francisco ZINC database.” Capturado em: http://zinc.docking.org/, Setembro 2009.

APÊNDICE A

Lotes gerados pelos testes experimentais com o algoritmo Hierarchical e com o ligante PIF

Este item apresenta os lotes gerados após aplicação do algoritmo para o P-MIA. Os *clusters* utilizados possuem diferentes quantidades, conforme Tabela 1.

Tabela 1 – Quantidade de *snapshots* em cada *cluster*

Cluster	**Quantidade**
*Cluster*_0	733
*Cluster*_1	822
*Cluster*_2	109
*Cluster*_3	107
*Cluster*_4	608
*Cluster*_5	663

As Tabelas 2, 3, 4, 5, 6 e 7 apresentam os lotes formados para os diferentes *clusters*.

Tabela 2 – Lotes para processamento do *cluster* 0

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
0	0	220	188 197 207 211 212 214 215 216 217 219 227 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 437 438
0	1	154	439 440 441 442 444 445 446 447 448 449 450 451 452 453 455 456 457 458 459 460 461 462 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 482 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 503 504 505 506 507 508 509 510 511 512 513 514 515 516 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 564 565 566 567 568 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 596 597 598 599 600 601 602
0	2	108	603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 647 648

			649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 692 693 696 697 698 699 700 701 702 703 705 706 707 708 709 710 711 712 713 714 715 715
0	3	76	716 717 719 720 721 722 724 726 727 728 730 733 734 735 737 739 740 741 743 744 745 746 748 749 750 751 752 755 756 758 759 760 761 762 764 765 766 767 768 769 770 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 801 802 803 804 805 806 808 808
0	4	53	809 810 811 813 814 815 816 817 818 819 820 821 823 825 831 833 837 846 847 849 862 864 867 868 869 871 872 873 876 877 879 880 881 882 883 887 888 889 892 893 895 896 898 899 901 902 904 905 906 908 909 913 917 917
0	5	50	919 921 923 924 926 927 929 930 933 935 936 938 939 940 949 955 958 959 960 962 968 971 972 978 983 990 991 992 994 995 996 997 998 1000 1001 1002 1003 1004 1005 1006 1008 1009 1011 1012 1014 1016 1017 1018 1019 1020 1020
0	6	72	1021 1026 1039 1042 1057 1059 1061 1062 1066 1068 1079 1080 1087 1088 1089 1091 1092 1094 1095 1096 1097 1098 1100 1101 1102 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1117 1118 1119 1120 1130 1131 1133 1134 1138 1142 1143 1155 1159 1164 1165 1166 1167 1168 1169 1170 1174 1186 1272 1278 1283 1284 1285 1286 1287 1289 1302 1303 1306 1309 1311 1312 1312

Tabela 3 – Lotes para processamento do *cluster* 1

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
1	0	247	481 483 704 812 824 826 827 828 829 830 832 834 835 836 838 839 841 842 843 844 845 848 850 852 853 855 858 859 860 861 863 865 866 870 875 878 885 890 891 894 897 900 903 907 910 911 912 914 915 916 918 920 922 925 928 931 932 934 937 941 942 943 944 945 946 947 948 950 951 952 953 954 956 957 961 963 964 965 966 967 969 970 973 974 975 976 977 979 980 981 982 984 985 986 987 988 989 993 999 1007 1010 1013 1015 1022 1023 1024 1025 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1040 1041 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1058 1060 1063 1064 1065 1067 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1081 1082 1083 1084 1085 1086 1090 1093 1099 1103 1104 1116 1121 1122 1123 1124 1125 1126 1127 1128 1129 1132 1135 1136 1137 1139 1140 1141 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1156 1157 1158 1160 1161 1162 1163 1171 1172 1173 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223
1	1	172	1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1273 1274 1275 1276 1277 1279 1280 1281 1282 1288 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1304 1305 1307 1308 1310 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410

1	2	121	1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532
1	3	85	1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1545 1546 1547 1548 1549 1550 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1563 1564 1566 1568 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1580 1581 1584 1585 1586 1587 1588 1589 1590 1591 1593 1594 1600 1601 1602 1603 1604 1605 1610 1622 1623 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1641 1642 1643 1644 1645 1647 1651 1652 1653 1654 1656
1	4	59	1657 1658 1660 1661 1662 1665 1668 1669 1673 1684 1685 1707 1710 1742 1746 1753 1757 1758 1759 1760 1761 1781 1782 1783 1792 1794 1801 1802 1803 1804 1806 1807 1813 1814 1819 1820 1821 1822 1824 1825 1826 1827 1828 1829 1830 1832 1839 1841 1843 1844 1845 1847 1848 1857 1858 1859 1861 1865 1867
1	5	50	1871 1873 1874 1875 1876 1877 1881 1884 1885 1886 1887 1891 1892 1893 1898 1902 1904 1906 1909 1911 1914 1916 1940 1959 1965 1968 1971 1972 1973 1974 1975 1976 1977 1978 1981 1982 1983 1984 1985 1986 1987 1988 1989 2000 2033 2041 2056 2057 2059 2124
1	6	88	2130 2139 2143 2179 2181 2182 2188 2196 2206 2287 2290 2298 2307 2311 2312 2314 2317 2320 2326 2327 2328 2329 2331 2338 2339 2340 2341 2343 2347 2348 2349 2350 2352 2353 2354 2355 2356 2357 2358 2360 2361 2364 2365 2368 2369 2373 2375 2377 2382 2383 2384 2385 2386 2388 2389 2392 2393 2418 2419 2420 2422 2423 2437 2469 2470 2471 2486 2487 2490 2491 2509 2512 2513 2514 2517 2519 2520 2521 2524 2525 2526 2529 2532 2533 2534 2535 2537 2548

Tabela 4 – Lotes para processamento do *cluster* 2

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
2	0	50	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
2	1	59	51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 97 98 99 100 101 104 105 106 107 148 151 152 153 158 158

Tabela 5 – Lotes para processamento do *cluster* 3

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
3	0	50	96 102 103 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 149 150 154 155 156 157 159
3	1	57	160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 189 190 191 192 193 194 195 196 198 199 200 201 202 203 204 205 206 208 209 210 213 218 221 222 223 224 225 226 228

Tabela 6 – Lotes para processamento do *cluster* 4

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
4	0	182	1200 1382 1431 1544 1551 1562 1565 1567 1582 1583 1592 1598 1599 1606 1607 1608 1609 1611 1612 1614 1615 1617 1618 1620 1621 1624 1626 1638 1639 1640 1649 1650 1659 1663 1664 1666 1667 1670 1671 1672 1675 1676 1677 1678 1679 1680 1681 1682 1683 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695 1696 1697 1699 1700 1701 1702 1703 1704 1705 1706 1708 1709 1711 1712 1713 1714 1715 1717 1718 1719 1720 1721 1722 1723 1724 1725 1726 1727 1728 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1740 1741 1743 1744 1745 1747 1748 1749 1750 1751 1752 1754 1755 1756 1762 1763 1764 1765 1766 1767 1768 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1784 1785 1786 1787 1788 1789 1790 1791 1793 1795 1796 1797 1798 1799 1800 1805 1809 1811 1812 1815 1816 1817 1818 1823 1831 1835 1838 1842 1846 1855 1856 1866 1868 1869 1872 1878 1879 1880 1882 1883 1888 1889 1890 1894 1895 1896 1897 1899 1900 1903
4	1	128	1905 1908 1910 1912 1913 1917 1918 1919 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1960 1961 1962 1963 1964 1966 1967 1969 1970 1979 1980 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2034 2035 2036 2037 2038 2039 2040 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2058 2060 2061 2062 2063 2064 2065 2066 2067
4	2	89	2068 2069 2070 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2123 2125 2126 2127 2128 2129 2131 2132 2133 2134 2135 2136 2137 2138 2140 2141 2142 2144 2145 2146 2147 2148 2149 2150 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160
4	3	63	2161 2163 2164 2165 2166 2167 2168 2170 2171 2172 2173 2174 2175 2176 2177 2178 2180 2183 2184 2185 2186 2187 2189 2190 2191 2192 2193 2194 2195 2197 2198 2199 2200 2201 2202 2203 2204 2205 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2224 2225 2226 2228 2229 2230 2232 2234 2240
4	4	50	2243 2245 2246 2248 2252 2253 2254 2255 2262 2263 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2289 2291 2293 2294 2295 2296 2297 2299 2301 2302 2305 2306 2308 2310 2313 2315 2316 2318 2319 2321 2323 2324 2325 2330
4	5	96	2335 2336 2337 2342 2344 2346 2351 2359 2362 2372 2374 2376 2381 2387 2394 2396 2397 2398 2400 2402 2432 2433 2444 2446 2448 2449 2450 2453 2454 2462 2476 2478 2479 2484 2488 2494 2499 2503 2505 2507 2527 2528 2530 2536 2562 2563 2564 2566 2573 2574 2578 2579 2612 2616 2626 2649 2651 2659 2663 2664 2665 2666 2669 2672 2675 2676 2677 2682 2685 2687 2688 2689 2690 2696 2697 2698 2699 2700 2701 2702 2703 2710 2711 2715 2723 2724 2727 2734 3075 3076 3079 3080 3081 3082 3089 3093

Tabela 7 – Lotes para processamento do *cluster* 5

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
5	0	199	1595 1596 1597 1613 1616 1619 1625 1646 1648 1655 1674 1698 1716 1808 1810 1833 1834 1836 1837 1840 1849 1850 1851 1852 1853 1854 1860 1862 1863 1864 1870 1901 1907 1915 1920 2162 2169 2223 2227 2231 2233 2235 2236 2237 2238 2239 2241 2242 2244 2247 2249 2250 2251 2256 2257 2258 2259 2260 2261 2264 2265 2266 2267 2268 2269 2270 2288 2292 2300 2303 2304 2309 2322 2332 2333 2334 2345 2363 2366 2367 2370 2371 2378 2379 2380 2390 2391 2395 2399 2401 2403 2404 2405 2406 2407 2408 2409 2410 2411 2412 2413 2414 2415 2416 2417 2421 2424 2425 2426 2427 2428 2429 2430 2431 2434 2435 2436 2438 2439 2440 2441 2442 2443 2445 2447 2451 2452 2455 2456 2457 2458 2459 2460 2461 2463 2464 2465 2466 2467 2468 2472 2473 2474 2475 2477 2480 2481 2482 2483 2485 2489 2492 2493 2495 2496 2497 2498 2500 2501 2502 2504 2506 2508 2510 2511 2515 2516 2518 2522 2523 2531 2538 2539 2540 2541 2542 2543 2544 2545 2546 2547 2549 2550 2551 2552 2553 2554 2555 2556 2557 2558 2559 2560 2561 2565 2567 2568 2569 2570
5	1	139	2571 2572 2575 2576 2577 2580 2581 2582 2583 2584 2585 2586 2587 2588 2589 2590 2591 2592 2593 2594 2595 2596 2597 2598 2599 2600 2601 2602 2603 2604 2605 2606 2607 2608 2609 2610 2611 2613 2614 2615 2617 2618 2619 2620 2621 2622 2623 2624 2625 2627 2628 2629 2630 2631 2632 2633 2634 2635 2637 2638 2639 2640 2641 2642 2643 2644 2645 2646 2647 2648 2650 2652 2653 2654 2655 2656 2657 2658 2660 2661 2662 2667 2668 2670 2671 2673 2674 2678 2679 2680 2681 2683 2691 2692 2693 2694 2695 2704 2705 2706 2707 2708 2709 2712 2713 2714 2716 2717 2718 2719 2720 2721 2722 2725 2726 2728 2729 2730 2731 2732 2733 2735 2736 2737 2738 2739 2740 2741 2742 2743 2744 2745 2746 2747 2748 2749 2750 2751 2753
5	2	97	2754 2755 2756 2757 2758 2759 2760 2761 2762 2763 2764 2765 2766 2767 2768 2769 2770 2771 2772 2774 2775 2776 2777 2778 2779 2780 2781 2782 2783 2784 2785 2786 2787 2788 2789 2790 2791 2792 2793 2794 2795 2796 2797 2798 2799 2800 2801 2802 2803 2804 2805 2806 2807 2808 2809 2810 2811 2812 2813 2814 2815 2816 2817 2818 2819 2820 2821 2822 2823 2824 2825 2826 2827 2828 2829 2830 2831 2832 2833 2834 2835 2836 2837 2838 2839 2840 2841 2842 2843 2844 2845 2846 2847 2848 2849 2850 2851
5	3	68	2852 2853 2854 2855 2856 2857 2858 2859 2860 2861 2862 2863 2864 2865 2866 2867 2868 2869 2870 2871 2872 2873 2874 2875 2876 2877 2878 2879 2880 2881 2882 2884 2885 2886 2887 2888 2889 2890 2891 2892 2893 2894 2895 2896 2897 2898 2899 2900 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 2911 2912 2913 2914 2915 2916 2917 2919 2920 2921
5	4	50	2922 2923 2924 2925 2926 2927 2928 2929 2930 2931 2932 2933 2934 2935 2936 2937 2938 2939 2940 2941 2942 2943 2944 2945 2946 2947 2948 2949 2950 2951 2952 2953 2954 2955 2956 2957 2958 2959 2960 2961 2962 2963 2964 2965 2966 2967 2968 2969 2970 2971
5	5	50	2975 2976 2978 2979 2980 2984 2985 2986 2988 2989 2990 2993 2994 2995 2996 2998 2999 3000 3001 3002 3003 3004 3005 3006 3007 3008 3009 3010 3011 3012 3013 3014 3015 3016 3017 3018 3019 3020 3021 3022 3023 3024 3025 3026 3027 3028 3029 3030 3031 3032
5	6	60	3033 3034 3035 3036 3037 3038 3039 3040 3041 3042 3043 3044 3045 3046 3047 3048 3049 3050 3051 3052 3053 3054 3055 3056 3057 3058 3059 3060 3061 3062 3063 3064 3065 3066 3067 3068 3069 3070 3071 3072 3073 3074 3077 3078 3083 3084 3085 3086 3087 3088 3090 3091 3092 3094 3095 3096 3097 3098 3099 3100 3100

Lotes gerados pelos testes experimentais com o algoritmo K-Means e com o ligante PIF

Os *clusters* utilizados possuem diferentes quantidades, conforme Tabela 8.

Tabela 8 – Quantidade de *snapshots* em cada *cluster* – agrupados conforme algoritmo *Means*

Cluster	**Quantidade**
*Cluster*_0	144
*Cluster*_1	552
*Cluster*_2	484
*Cluster*_3	140
*Cluster*_4	529
*Cluster*_5	1193

As Tabelas 9, 10, 11, 12, 13 e 14 apresentam os lotes formados para os diferentes *clusters*.

Tabela 9 – Lotes para processamento do *cluster* 0 - Means

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
0	0	50	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
0	1	94	51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 112 116 117 118 119 120 121 122 123 126 127 129 132 133 134 136 137 138 140 142 144 147 148 149 151 152 153 154 155 157 158 159 160 161 162 165

Tabela 10 – Lotes para processamento do *cluster* 1 - Means

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
1	0	166	1595 1596 1597 1613 1619 1681 1692 1698 1714 1715 1716 1717 1726 1732 1768 1836 1837 1852 2018 2073 2108 2123 2152 2158 2159 2162 2164 2166 2167 2168 2223 2227 2233 2235 2237 2239 2241 2242 2244 2249 2250 2251 2254 2255 2256 2257 2258 2259 2260 2264 2265 2266 2267 2268 2300 2405 2407 2408 2410 2411 2413 2414 2429 2440 2441 2451 2455 2458 2459 2504 2515 2522 2543 2545 2546 2549 2550 2551 2552 2556 2557 2558 2559 2565 2568 2571 2572 2577 2580 2582 2583 2585 2586 2588 2589 2590 2591 2592 2593 2594 2595 2596 2598 2599 2600 2602 2603 2604 2605 2607 2608 2609 2610 2613 2614 2615 2617 2618 2619 2620 2621 2622 2623 2624 2625 2626 2627 2628 2629 2630 2631 2632 2633 2634 2635 2637 2638 2639 2640 2641 2642 2643 2644 2645 2646 2647 2648 2649 2650 2652 2653 2654 2655 2656 2657 2658 2659 2660 2661 2662 2663 2664 2666 2667 2668 2669
1	1	116	2670 2671 2672 2673 2674 2676 2677 2678 2679 2680 2681 2682 2683 2690 2691 2692 2693 2694 2695 2696 2697 2703 2704 2705 2706 2707 2708 2709 2712 2713 2714 2715 2716 2717 2718 2719 2720 2721 2722 2723 2725 2726 2727 2728 2729 2730 2731 2732 2733 2734 2735 2736 2737 2738 2739 2740 2741 2743 2744 2745 2746 2747 2748 2749 2750 2751 2753 2754 2755 2756 2757 2758 2759 2760 2761 2762 2763 2764 2765 2766 2767 2768 2769 2770 2771 2772 2774 2775 2776 2777 2778 2779 2780 2781 2782 2783 2784 2785 2786 2787 2788 2789 2790 2791 2792 2793 2794 2795 2796 2797 2798 2799 2801 2802 2803 2804
1	2	81	2805 2806 2807 2808 2809 2810 2811 2812 2813 2814 2815 2816 2817 2818 2819 2820 2821 2822 2823 2824 2825 2826 2827 2828 2829 2830 2831 2832 2833 2834 2835 2836 2837 2838 2839 2840 2841 2842 2843 2844 2845 2846 2847 2848 2849 2850 2851 2852

			2853 2854 2855 2856 2857 2858 2859 2860 2861 2862 2863 2864 2865 2866 2867 2868 2869 2870 2871 2872 2873 2874 2875 2876 2877 2878 2879 2880 2881 2882 2884 2885 2886
1	3	57	2887 2888 2889 2890 2891 2892 2893 2894 2895 2896 2897 2898 2899 2900 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 2911 2912 2913 2914 2915 2916 2917 2919 2920 2921 2922 2923 2924 2925 2926 2927 2928 2929 2930 2931 2932 2933 2934 2935 2936 2937 2938 2939 2940 2941 2942 2943 2944
1	4	50	2945 2946 2947 2948 2949 2950 2951 2952 2953 2954 2955 2956 2957 2958 2959 2960 2961 2962 2963 2964 2965 2966 2967 2968 2969 2970 2971 2975 2976 2978 2979 2980 2984 2985 2986 2988 2989 2990 2993 2994 2995 2996 2998 2999 3000 3001 3002 3003 3004 3005
1	5	82	3006 3007 3008 3009 3010 3011 3012 3013 3014 3015 3016 3017 3018 3019 3020 3021 3022 3023 3024 3025 3026 3027 3028 3030 3031 3032 3033 3034 3035 3036 3037 3038 3039 3040 3041 3042 3043 3044 3045 3046 3047 3048 3049 3050 3051 3052 3053 3054 3056 3057 3058 3059 3060 3061 3062 3063 3064 3065 3066 3068 3069 3070 3071 3072 3073 3074 3075 3076 3077 3086 3088 3090 3091 3092 3093 3094 3095 3096 3097 3098 3099 3100

Tabela 11 – Lotes para processamento do *cluster* 2 - Means

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
2	0	145	273 277 287 288 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430
2	1	102	431 432 433 434 435 437 438 439 440 441 442 444 445 446 447 448 449 450 451 452 453 455 456 457 458 459 460 461 462 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 483 484 485 486 487 488 489 490 491 492 493 494 495 496 500 501 503 504 505 506 508 509 510 511 512 513 514 515 516 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544
2	2	71	545 546 547 548 549 550 551 552 553 554 555 556 557 559 560 561 562 564 565 566 567 568 570 571 572 573 574 575 576 577 578 580 582 583 584 585 586 587 588 589 590 591 592 593 594 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621
2	3	50	622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672
2	4	50	673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 692 693 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 717 721 724 727 728 730 734 735 737 739
2	5	66	741 743 745 749 750 752 755 756 758 759 760 761 762 764 765 766 767 768 769 770 772 773 774 775 776 777 779 780 781 782 783 784 786 787 788 789 790 791 792 793 794 795 796 798 799 801 802 803 804 805 809 810 823 862 864 869 876 879 880 882 883 885 887 888 896 1068

Tabela 12 – Lotes para processamento do *cluster* 3 - Means

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
3	0	50	109 110 111 113 114 115 124 125 128 130 131 135 139 141 143 145 146 150 156 163 164 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194
3	1	90	195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 274 275 276 278 279 280 281 282 283 284 285 286 289

Tabela 13 – Lotes para processamento do *cluster* 4 - Means

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
4	0	159	481 482 497 498 499 507 558 579 581 716 719 720 722 726 733 740 744 746 748 751 778 785 797 806 808 811 812 813 814 815 816 817 818 819 820 821 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 841 843 844 845 846 847 848 849 850 852 853 855 858 859 861 863 865 866 867 868 870 871 872 873 875 877 878 881 889 890 891 892 893 894 895 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968
4	1	111	969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080
4	2	78	1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1152 1153 1154 1155 1156 1157 1158 1159
4	3	54	1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1203 1204 1205 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220
4	4	50	1221 1222 1223 1224 1226 1227 1228 1229 1231 1232 1233 1234 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273
4	5	77	1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1326 1328 1329 1331 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1359 1361 1385 1386 1461

Tabela 14 – Lotes para processamento do *cluster* 5 - Means

Cluster	Lote	Quantidade de *Snapshots*	*Snapshots* do Lote
5	0	356	842 860 1151 1198 1199 1200 1201 1202 1206 1207 1225 1230 1235 1325 1327 1330 1332 1333 1334 1335 1336 1337 1338 1339 1357 1358 1360 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1382 1383 1384 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590 1591 1592 1593 1594 1598 1599 1600 1601 1602 1603 1604 1605 1606 1607 1608 1609 1610 1611 1612 1614 1615 1616 1617 1618 1620 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1693 1694 1695 1696 1697 1699 1700 1701
5	1	251	1702 1703 1704 1705 1706 1707 1708 1709 1710 1711 1712 1713 1718 1719 1720 1721 1722 1723 1724 1725 1727 1728 1729 1730 1731 1733 1734 1735 1736 1737 1738 1739 1740 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830 1831 1832 1833 1834 1835 1838 1839 1840 1841 1842 1843 1844 1845 1846 1847 1848 1849 1850 1851 1853 1854 1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962
5	2	176	1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071

			2072 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2124 2125 2126 2127 2128 2129 2130 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142
5	3	123	2143 2144 2145 2146 2147 2148 2149 2150 2151 2153 2154 2155 2156 2157 2160 2161 2163 2165 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2224 2225 2226 2228 2229 2230 2231 2232 2234 2236 2238 2240 2243 2245 2246 2247 2248 2252 2253 2261 2262 2263 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297
5	4	86	2298 2299 2301 2302 2303 2304 2305 2306 2307 2308 2309 2310 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358 2359 2360 2361 2362 2363 2364 2365 2366 2367 2368 2369 2370 2371 2372 2373 2374 2375 2376 2377 2378 2379 2380 2381 2382 2383 2384
5	5	60	2385 2386 2387 2388 2389 2390 2391 2392 2393 2394 2395 2396 2397 2398 2399 2400 2401 2402 2403 2404 2406 2409 2412 2415 2416 2417 2418 2419 2420 2421 2422 2423 2424 2425 2426 2427 2428 2430 2431 2432 2433 2434 2435 2436 2437 2438 2439 2442 2443 2444 2445 2446 2447 2448 2449 2450 2452 2453 2454 2456
5	6	50	2457 2460 2461 2462 2463 2464 2465 2466 2467 2468 2469 2470 2471 2472 2473 2474 2475 2476 2477 2478 2479 2480 2481 2482 2483 2484 2485 2486 2487 2488 2489 2490 2491 2492 2493 2494 2495 2496 2497 2498 2499 2500 2501 2502 2503 2505 2506 2507 2508 2509
5	7	91	2510 2511 2512 2513 2514 2516 2517 2518 2519 2520 2521 2523 2524 2525 2526 2527 2528 2529 2530 2531 2532 2533 2534 2535 2536 2537 2538 2539 2540 2541 2542 2544 2547 2548 2553 2554 2555 2560 2561 2562 2563 2564 2566 2567 2569 2570 2573 2574 2575 2576 2578 2579 2581 2584 2587 2597 2601 2606 2611 2612 2616 2651 2665 2675 2685 2687 2688 2689 2698 2699 2700 2701 2702 2710 2711 2724 2742 2800 3029 3055 3067 3078 3079 3080 3081 3082 3083 3084 3085 3087 3089

Lotes gerados pelos testes experimentais com o algoritmo K-Means e com o ligante NADH

As Tabelas 15, 16, 17, 18, 19 e 20 apresentam os lotes formados para os diferentes *clusters*.

Tabela 15 – Lotes para processamento do *cluster* 0 – Means com NADH

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
0	0	50	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
0	1	94	51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 112 116 117 118 119 120 121 122 123 126 127 129 132 133 134 136 137 138 140 142 144 147 148 149 151 152 153 154 155 157 158 159 160 161 162 165

Tabela 16 – Lotes para processamento do *cluster* 1 – Means com NADH

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
1	0	171	1595 1596 1597 1613 1619 1681 1692 1698 1714 1715 1716 1717 1726 1732 1768 1836 1837 1852 2018 2073 2108 2123 2152 2158 2159 2162 2164 2166 2167 2168 2223 2227 2233 2235 2237 2239 2241 2242 2244 2249 2250 2251 2254 2255 2256 2257 2258 2259 2260 2264 2265 2266 2267 2268 2300 2405 2407 2408 2410 2411 2413 2414 2429 2440 2441 2451 2455 2458 2459 2504 2515 2522 2543 2545 2546 2549 2550 2551 2552 2556 2557 2558 2559 2565 2568 2571 2572 2577 2580 2582 2583 2585 2586 2588 2589 2590 2591 2592 2593 2594 2595 2596 2598 2599 2600 2602 2603 2604 2605 2607 2608 2609 2610 2613 2614 2615 2617 2618 2619 2620 2621 2622 2623 2624 2625 2626 2627 2628 2629 2630 2631 2632 2633 2634 2635 2636 2637 2638 2639 2640 2641 2642 2643 2644 2645 2646 2647 2648 2649 2650 2652 2653 2654 2655 2656 2657 2658 2659 2660 2661 2662 2663 2664 2666 2667 2668 2669 2670 2671 2672 2673
1	1	120	2674 2676 2677 2678 2679 2680 2681 2682 2683 2684 2690 2691 2692 2693 2694 2695 2696 2697 2703 2704 2705 2706 2707 2708 2709 2712 2713 2714 2715 2716 2717 2718 2719 2720 2721 2722 2723 2725 2726 2727 2728 2729 2730 2731 2732 2733 2734 2735 2736 2737 2738 2739 2740 2741 2743 2744 2745 2746 2747 2748 2749 2750 2751 2752 2753 2754 2755 2756 2757 2758 2759 2760 2761 2762 2763 2764 2765 2766 2767 2768 2769 2770 2771 2772 2773 2774 2775 2776 2777 2778 2779 2780 2781 2782 2783 2784 2785 2786 2787 2788 2789 2790 2791 2792 2793 2794 2795 2796 2797 2798 2799 2801 2802 2803 2804 2805 2806 2807 2808 2809
1	2	83	2810 2811 2812 2813 2814 2815 2816 2817 2818 2819 2820 2821 2822 2823 2824 2825 2826 2827 2828 2829 2830 2831 2832 2833 2834 2835 2836 2837 2838 2839 2840 2841 2842 2843 2844 2845 2846 2847 2848 2849 2850 2851 2852 2853 2854 2855 2856 2857 2858 2859 2860 2861 2862 2863 2864 2865 2866 2867 2868 2869 2870 2871 2872 2873 2874 2875 2876 2877 2878 2879 2880 2881 2882 2883 2884 2885 2886 2887 2888 2889 2890 2891 2892
1	3	58	2893 2894 2895 2896 2897 2898 2899 2900 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 2911 2912 2913 2914 2915 2916 2917 2918 2919 2920 2921 2922 2923 2924 2925 2926 2927 2928 2929 2930 2931 2932 2933 2934 2935 2936 2937 2938 2939 2940 2941 2942 2943 2944 2945 2946 2947 2948 2949 2950
1	4	50	2951 2952 2953 2954 2955 2956 2957 2958 2959 2960 2961 2962 2963 2964 2965 2966 2967 2968 2969 2970 2971 2972 2973 2974 2975 2976 2977 2978 2979 2980 2981 2982 2983 2984 2985 2986 2987 2988 2989 2990 2991 2992 2993 2994 2995 2996 2997 2998 2999 3000
1	5	87	3001 3002 3003 3004 3005 3006 3007 3008 3009 3010 3011 3012

			3013 3014 3015 3016 3017 3018 3019 3020 3021 3022 3023 3024 3025 3026 3027 3028 3030 3031 3032 3033 3034 3035 3036 3037 3038 3039 3040 3041 3042 3043 3044 3045 3046 3047 3048 3049 3050 3051 3052 3053 3054 3056 3057 3058 3059 3060 3061 3062 3063 3064 3065 3066 3068 3069 3070 3071 3072 3073 3074 3075 3076 3077 3086 3088 3090 3091 3092 3093 3094 3095 3096 3097 3098 3099 3100

Tabela 17 – Lotes para processamento do *cluster* 2 – Means com NADH

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
2	0	152	273 277 287 288 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437
2	1	107	438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 483 484 485 486 487 488 489 490 491 492 493 494 495 496 500 501 503 504 505 506 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551
2	2	75	552 553 554 555 556 557 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 580 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629
2	3	52	630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681
2	4	50	682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 717 718 721 724 727 728 730 731 734 735 736 737 739 741 742 743
2	5	72	745 747 749 750 752 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 772 773 774 775 776 777 779 780 781 782 783 784 786 787 788 789 790 791 792 793 794 795 796 798 799 800 801 802 803 804 805 809 810 823 862 864 869 874 876 879 880 882 883 884 885 886 887 888 896 1068

Tabela 18 – Lotes para processamento do *cluster* 3 – Means com NADH

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
3	0	50	109 110 111 113 114 115 124 125 128 130 131 135 139 141 143 145 146 150 156 163 164 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194
3	1	91	195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 274 275 276 278 279 280 281 282 283 284 285 286

			289

Tabela 19 – Lotes para processamento do *cluster* 4 – Means com NADH

Cluster	**Lote**	**Quantidade de *Snapshots***	***Snapshots* do Lote**
4	0	163	481 482 497 498 499 502 507 558 579 581 716 719 720 722 723 725 726 729 732 733 738 740 744 746 748 751 753 771 778 785 797 806 807 808 811 812 813 814 815 816 817 818 819 820 821 822 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 861 863 865 866 867 868 870 871 872 873 875 877 878 881 889 890 891 892 893 894 895 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957
4	1	114	958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1069 1070 1071 1072
4	2	80	1073 1074 1075 1076 1077 1078 1079 1080 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1152 1153
4	3	56	1154 1155 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1203 1204 1205 1208 1209 1210 1211 1212 1213 1214 1215 1216
4	4	50	1217 1218 1219 1220 1221 1222 1223 1224 1226 1227 1228 1229 1231 1232 1233 1234 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269
4	5	81	1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1326 1328 1329 1331 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1359 1361 1385 1386 1461

Tabela 20 – Lotes para processamento do *cluster* 5 – Means com NADH

Cluster	Lote	Quantidade de *Snapshots*	*Snapshots* do Lote
5	0	358	842 860 1151 1198 1199 1200 1201 1202 1206 1207 1225 1230 1235 1325 1327 1330 1332 1333 1334 1335 1336 1337 1338 1339 1357 1358 1360 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1382 1383 1384 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590 1591 1592 1593 1594 1598 1599 1600 1601 1602 1603 1604 1605 1606 1607 1608 1609 1610 1611 1612 1614 1615 1616 1617 1618 1620 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1693 1694 1695 1696 1697 1699 1700 1701 1702 1703
5	1	251	1704 1705 1706 1707 1708 1709 1710 1711 1712 1713 1718 1719 1720 1721 1722 1723 1724 1725 1727 1728 1729 1730 1731 1733 1734 1735 1736 1737 1738 1739 1740 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830 1831 1832 1833 1834 1835 1838 1839 1840 1841 1842 1843 1844 1845 1846 1847 1848 1849 1850 1851 1853 1854 1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964
5	2	176	1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071 2072 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2124

			2125 2126 2127 2128 2129 2130 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144
5	3	123	2145 2146 2147 2148 2149 2150 2151 2153 2154 2155 2156 2157 2160 2161 2163 2165 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2224 2225 2226 2228 2229 2230 2231 2232 2234 2236 2238 2240 2243 2245 2246 2247 2248 2252 2253 2261 2262 2263 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297 2298 2299
5	4	86	2301 2302 2303 2304 2305 2306 2307 2308 2309 2310 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358 2359 2360 2361 2362 2363 2364 2365 2366 2367 2368 2369 2370 2371 2372 2373 2374 2375 2376 2377 2378 2379 2380 2381 2382 2383 2384 2385 2386
5	5	60	2387 2388 2389 2390 2391 2392 2393 2394 2395 2396 2397 2398 2399 2400 2401 2402 2403 2404 2406 2409 2412 2415 2416 2417 2418 2419 2420 2421 2422 2423 2424 2425 2426 2427 2428 2430 2431 2432 2433 2434 2435 2436 2437 2438 2439 2442 2443 2444 2445 2446 2447 2448 2449 2450 2452 2453 2454 2456 2457 2460
5	6	50	2461 2462 2463 2464 2465 2466 2467 2468 2469 2470 2471 2472 2473 2474 2475 2476 2477 2478 2479 2480 2481 2482 2483 2484 2485 2486 2487 2488 2489 2490 2491 2492 2493 2494 2495 2496 2497 2498 2499 2500 2501 2502 2503 2505 2506 2507 2508 2509 2510 2511
5	7	90	2512 2513 2514 2516 2517 2518 2519 2520 2521 2523 2524 2525 2526 2527 2528 2529 2530 2531 2532 2533 2534 2535 2536 2537 2538 2539 2540 2541 2542 2544 2547 2548 2553 2554 2555 2560 2561 2562 2563 2564 2566 2567 2569 2570 2573 2574 2575 2576 2578 2579 2581 2584 2587 2597 2601 2606 2611 2612 2616 2651 2665 2675 2685 2686 2687 2688 2689 2698 2699 2700 2701 2702 2710 2711 2724 2742 2800 3029 3055 3067 3078 3079 3080 3081 3082 3083 3084 3085 3087 3089