Qualitative knowledge models in Functional Genomics and Proteomics

Mor Peleg[1], Irene S. Gabashvili[2], and Russ B. Altman[3*]

[1]*Department of Management Information Systems, University of Haifa, Israel, 31905, +972-4-8288504*

morpeleg@mis.hevra.haifa.ac.il

[2]*Hewlett Packard Labs, MS 1169*

1501 Page Mill Road, Palo Alto, CA 94304

irene.gabashvili@hp.com

[3]*Stanford Medical Informatics, Stanford University, MSOB x-215,*

251 Campus Drive, Stanford, CA, 94305, USA

russ.altman@stanford.edu

[*] To whom correspondence should be addressed

1. Introduction

Predicting pathological phenotypes based on genetic mutations remains a fundamental and unsolved issue. When a gene is mutated, the molecular functionality of the gene product may be affected and many cellular processes may go awry. Basic molecular functions occur in networks of interactions and events that produce subsequent cellular and physiological functions. Most knowledge of these interactions is represented diffusely in the published literature, Excel lists and specialized relational databases and so it is difficult to assess our state of understanding at any moment. Thus it would be very useful to systematically store knowledge in data structures that allow the knowledge to be evaluated and examined in detail by scientists as well as computer algorithms. Our goal is to develop technology for representing qualitative, noisy, and sparse biological results in support of the eventual goal of fully accurate quantitative models.

In a recent paper, we described an ontology that we developed for modeling biological processes [1]. **Ontologies** provide consistent definitions and interpretations of concepts in a domain of interest (e.g., biology), and enable software applications to share and reuse the knowledge consistently [2]. Ontologies can be used to perform logical inference over the set of concepts to provide for generalization and explanation facilities [3]. Our biological process ontology combines and extends two existing components: a Workflow model and a biomedical-ontology, both described in the Methods and Tools section. Our resulting framework possesses the following properties: (1) it allows qualitative modeling of structural and functional aspects of a biological system, (2) it includes biological and medical concept models to allow for querying biomedical information using biomedical abstractions, (3) it allows hierarchical models to manage the complexity of the representation, (4) it has a sound logical basis for automatic verification, and (5) it has an intuitive, graphical representation.

Our application domain is tRNA-related disease. tRNA constitutes a good test-bed because there exists rich literature on tRNA molecular structure as well as the diseases that result from abnormal structures in mitochondria (many of which affect neural processes) The main role of tRNA molecules is to be part of the machinery for the translation of the genetic message, encoded in mRNA, into a protein. This process employs over twenty different tRNA molecules, each specific for one amino acid and for a particular triplet of nucleotides in mRNA (codon) [4]. Several steps take place before a tRNA molecule can participate in translation. After a gene coding for tRNA is transcribed, the RNA product is folded and processed to become a tRNA molecule. tRNA molecules are covalently linked (acylated) with an amino acid to form amino-acylated-tRNA (aa-tRNA). The aa-tRNA molecules can then bind with translation factors to form complexes that may participate in the translation process. There are three kinds of complexes that participate in translation: (i) an *initiation complex* is formed by exhibiting tRNA mimicry release factors that bind to the stop codon in the mRNA template or a by mis-functioning tRNA complexed with GTP and elongation factor causing abnormal termination, and (iii) a *ternary complex* is formed by binding Elongating aa-tRNAs (tRNAs that are acylated to amino acids other than formylmethionine) with GTP and the elongation factor EF-tu. During the translation process, tRNA molecules recognize the mRNA codons one by one, as the mRNA molecule moves through the cellular machine for protein synthesis: the ribosome. In 1964, Watson introduced the classical two-site model, which was the accepted model until 1984 [5]. In this model, the ribosome has two regions for tRNA binding, so-called aminoacyl (A)-site and peptidyl (P)-site. According to this model, initiation starts from the P-site, but during the normal cycle of elongation, each tRNA enters the ribosome from the A- site and proceeds to the P-site before exiting into the cell's cytoplasm. Currently, it is hypothesized that the ribosome has at

least three regions for tRNA binding: the A and P sites, and an exit site (E-site) through which the tRNA exits the ribosome into the cell's cytoplasm [6]. Protein synthesis is terminated when a stop codon is reached at the ribosomal A-site and recognized by a specific termination complex, probably involving factors mimicking tRNA. Premature termination (e.g., due to a mutation in tRNA) can be also observed [7].

When aa-tRNA molecules bind to the A-site, they normally recognize and bind to matching mRNA codons – a process known as reading. tRNA mutations can cause abnormal reading that leads to mutated protein products of translation. Types of abnormal reading include: (1) *Misreading*, where tRNA with non-matching amino-acid binds to the ribosome's A-site, (2) *frame-shifting*, where tRNA that causes frame-shifting (e.g., binds to four nucleotides of the mRNA at the A-site) participates in elongation, and (3) *Halting*, where tRNA that cause premature termination (e.g., tRNA that are not acetylated with an amino acid) binds to the A-site. These three types of errors, along with the inability to bind to the A-site or destruction by cellular enzymes due to misfolding, can create complex changes in protein profiles of cells. This can affect all molecular partners of produced proteins in the chain of events connecting genotype to phenotype and produce a variety of phenotypes. Mutations in human tRNA molecules have been implicated in a wide range of disorders including myopathies, encephalopathies, cardiopathies, diabetes, growth retardation, and aging [8]. Development of models that consolidate and integrate our understanding of the molecular foundations for these diseases, based on available structural, biochemical and physiological knowledge, is therefore urgently needed.

In a recent paper [9], we discussed an application of our biological process ontology to genomics and proteomics. The current article extends the section on general computer-science theories, including Petri Nets, ontologies, and information systems modeling methodologies, as

well as extends the section on biological sources of information and discusses the compatibility of our outputs with popular databases and modeling environments.

The paper is organized as follows. Section 2 describes the components we used to develop the framework and the knowledge sources for our model. Section 3 discusses our modeling approach and demonstrates our knowledge model and the way in which information can be viewed and queried using the process of translation as examples. We conclude with a discussion and conclusion.

2. Methods and Tools

Component ontologies

Our framework combines and extends two existing components: Workflow model and biomedical ontology. The **Workflow model** [10] consists of a Process Model and an Organizational (Participants/Role) Model. The **Process Model** can represent ordering of processes (e.g., protein translation) and the structural components that participate in them (e.g., protein). Processes may be of low granularity (high-level processes) or of high granularity (low-level processes). High-level processes are nested to control the complexity of the presentation for human inspection. The **Participants/Role Model** represents the relationships among participants (e.g., an EF-tu is a member of the Elongation Factors collection in prokaryotes) and the roles that participants play in the modeled processes (e.g., EF-tu has enzymic function: GTPase). We used the workflow model as a biological process model by mapping Workflow activities to biological processes; organizational units to biomolecular complexes; Humans (individuals) to their biopolymers and networks of events; and roles to biological processes and functions.

A significant advantage of the workflow model is that it can map to **Petri Nets** [11], a mathematical model that represents concurrent systems, which allows verification of formal

properties as well as qualitative simulation [12]. A **Petri Net** is represented by a directed, bipartite graph in which nodes are either places or transitions, where places represent conditions (e.g., parasite in blood stream) and transitions represent activities (e.g., invasion of host erythrocytes). Tokens that are placed on places define the state of the Petri Net (marking). A token that resides in a place signifies that the condition that the place represents is true. A Petri Net can be executed in the following way. When all the places with arcs to a transition have a token, the transition is enabled, and may fire, by removing a token from each input place and adding a token to each place pointed to by the transition. **High-level Petri Nets,** used in this work, include extensions that allow modeling of time, data, and hierarchies.

For the biomedical ontology, we combine the Transparent Access to Multiple Biological Information Sources (TAMBIS) [13] with The Unified Medical Language System (UMLS) [14]. TAMBIS is an ontology for describing data to be obtained from bioinformatics sources. It describes biological entities at the molecular level. UMLS describes clinical and medical entities. It is a publicly available federation of biomedical controlled-terminologies and includes a Semantic Network with 134 semantic types that provide a consistent categorization of thousands of biomedical concepts. The 2002AA edition of the UMLS Metathesaurus includes 776,940 concepts and 2.1 million concept names in over 60 different biomedical source vocabularies. We augmented these two core terminological models [1] to represent mutations and their affects on biomolecular structures, biochemical functions, cellular processes, and clinical phenotypes. The extensions include classes for representing: (1) mutations and alleles, and their relationship to sequence components, (2) nucleic acid 3D structure linked to secondary and primary structural blocks, and (3) a set of composition operators, based on the nomenclature of composition relationships, due to Odell [15].

Odell introduced a nomenclature of six kinds of composition. We are using three of these composition relationships in our model. The relationship between a biomolecular complex (e.g., Ternary complex) and its parts (e.g., GTP, EF-tu, aa-tRNA) is a *component-integral object composition*. This relationship defines a *configuration* of parts within a whole. A *configuration* requires the parts to bear a particular functional or structural relationship to one another, as well as to the object they constitute. The relationship between an individual molecule (e.g., tRNA) and its domains (e.g., D-Domain, T-domain) is a *place-area composition*. This relationship defines a configuration of parts, where parts are the same kind of thing as the whole, and the parts cannot be separated from the whole. *Member-bunch composition* groups together molecules into collections when the collection members share similar functionality (e.g., elongation factors) or cellular location (e.g., membrane proteins). We have not found the other three composition relationships due to Odell to be relevant for our model.

We implemented our framework using the Protégé-2000 knowledge-modeling tool [16]. We used Protégé's axiom language (PAL) to define queries in a subset of first-order predicate logic written in the Knowledge Interchange Format syntax. The queries present, in tabular format, relationships among processes and structural components, as well as the relationship between a defective process or clinical phenotype to the mutation that is causing it.

Translation into Petri Nets

We manually translated the tRNA workflow model into corresponding Petri Nets, according to mapping defined by others [12]. The Petri Net models that we used were high-level Petri Nets that allow the representation of hierarchy and data. Hierarchies enable expanding a transition in a given Petri Net to an entire Petri Net, as is done in expanding Workflow high-level processes into a net of lower-level processes. We upgraded the derived Petri Nets to Colored Petri Nets (CPNs), by:

(1) Defining color sets for tRNA molecules (mutated and normal), mRNA molecules, and nucleotides that comprise the mRNA sequence, and initiating the Petri Nets with an initial marking of colored tokens

(2) Adding guards on transitions that relate to different types of tRNA molecules (e.g., fMet-tRNA vs. elongating tRNA molecules)

(3) Defining mRNA sequences that serve as the template for translation

We used the *Woflan* Petri Net verification tool [17] to verify that the Petri Nets are bounded (i.e., no accumulation of an infinite amount of tokens) and live (i.e., deadlocks do not exist). To accommodate limitations in the Woflan tool, which does not support colored Petri Nets, we manually made several minor changes to the Petri Nets before verifying them. We simulated the Petri Nets to study the dynamic aspects of the translation process, using the Design CPN tool [18], that has since been replaced by *CPN Tools*.

Sources of biological data

We gathered information from databases and published literature in order to develop the tRNA example considered in this work. We identified data sources with information pertaining to tRNA sequence, structure, modifications, mutations and disease associations. The databases that we used were:

- Compilation of mammalian mitochondrial tRNA genes [19], aimed at defining typical as well as consensus primary and secondary structural features of mammalian mitochondrial tRNAs (http://mamit-trna.u-strasbg.fr/)
- Compilation of tRNA sequences and sequences of tRNA genes [20] (http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/)

- The Comparative RNA Web site (http://www.rna.icmb.utexas.edu/), which provides a modeling environment for sequence and secondary structure comparisons [21]
- Structural Classifications Of RNA (SCOR, http://scor.lbl.gov/scor.html) [22]
- The RNA Modification Database (http://medlib.med.utah.edu/RNAmods), which provides literature and data on nucleotide modifications in RNA [23]
- A database on tRNA genes and molecules in mitochondria and photosynthetic eukaryotes (http://www.ba.itb.cnr.it/PLMItRNA/) [8]
- Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim/), which catalogues human genes and genetic disorders [24]
- BioCyc (http://metacyc.org/). BioCyc is a collection of genome and metabolic-pathway databases, which describes pathways, reactions, and enzymes of a variety of organisms [25].
- Entrez, the Life Sciences Search Engine, provides views for a variety of genomes, complete chromosomes, contiged sequence maps, and integrated genetic and physical maps (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi?itool=toolbar) [26].
- MITOMAP - A human mitochondrial genome database [27] (http://www.mitomap.org/)
- The UniProt/Swiss-Prot Protein Knowledgebase giving access to wealthy annotations and publicly available resources of protein information (http://us.expasy.org/sprot/sprot-top.html)

In addition, we used microarrays [28] and mass spectra data [29], providing information on proteins involved in tRNA processing or affected by tRNA mutations.

3. Modeling Approach and Results

Our model represents data using Process Diagrams and Participant-Role Diagrams. Appendix A on our web site (http://mis.hevra.haifa.ac.il/~morpeleg/NewProcessModel/Malaria_PN_Example_Files.html) presents the number of processes, participants, roles, and links that we used in our model. The most granular thing that we represented was at the level of a single nucleotide (e.g., GTP). The biggest molecule that we represented was the ribosome. We chose our levels of granularity in a way that considers the translation process under the assumption of a perfect ribosome; we only considered errors in translation that are due to tRNA. This assumption also influenced our design of the translation process model. This design follows individual tRNA molecules throughout the translation process, and therefore represents the translocation of tRNA molecules from the P to E site, and from the A to P site as distinct processes that occur in parallel. The level of detail in which we represented the model led us to consider questions such as: (1) “Can tRNA bind the A-site before previously bound tRNA molecule is released from the E-site?”, and (2) “Can fMet-tRNA form a ternary complex?”.

Representing mutations

Variation in gene products (proteins or RNA) can result from mutations in the nucleotide sequence of a gene, leading to altered (1) translation, (2) splicing, (3) post-transcriptional end-processing, or (4) interactions with other cellular components co-participating in biological processes. In addition, variation can result from a normal sequence that is translated improperly by abnormal tRNA molecules. Thus, we must be able to represent variation not only in DNA sequences (genome), but also in RNA and protein. Therefore, in our ontology, every *Sequence_Component* (of a nucleic acid or protein) may be associated with multiple *Alleles*.

Each *Allele* may have *Mutations* that are either pathogenic (associated with abnormal functions) or neutral. A mutation is classified as a substitution, insertion, or deletion [30].

Representing nucleic acid structure

The TAMBIS terminology did not focus on 3D structure. We extended the TAMBIS ontology by specifying tertiary structure components of nucleic acids. A nucleic acid tertiary structure component is composed of interacting segments of nucleic acid secondary structure components. We added three types of nucleic acid secondary structure components: nucleic acid helix, nucleic acid loop, and nucleic acid unpaired strand. Figure 1 shows the tertiary structure components of tRNA (Acceptor Domain, D-Domain, T-domain, Variable Loop, and Anticodon Domain). Also shown is the nucleic acid tertiary structure component frame that corresponds to the tRNA Acceptor Domain. The division of tRNA into structural domains, the numbering of nucleotides of the generic tRNA molecule, and the sequence-to-structure correspondence was done according to conventional rules [20].

Insert Figure 1 here

Representing molecular complexes

Biological function can be associated with different levels of molecular structure. In some cases a function can be associated with a domain (of a protein or nucleic acid). In other cases, a function is associated with individual molecules, or with molecular complexes. Some times, a function is not specifically mapped to a molecular structure, but is attributed to collections of

molecules that are located in a particular cellular compartment. In addition, biologists define collections of molecules that share a common function (e.g., termination factors). The Participant/Role representation of our framework represents molecular structures that participate in processes, as well as composition and generalization relationships among participants (molecules).

In our tRNA example, we are using three kinds of these composition relationships: (1) component-integral object composition, (2) member-bunch composition, and (3) place-area composition. Figure 2 shows examples of these relationships. Generalization (is-a) relationships are used to relate sub-classes of participants to their super-classes. For example, *Terminator tRNA*, *Non-terminating tRNA*, and *fMet tRNA* are sub-classes of *the tRNA class*.

Insert Figure 2 here

Representing abnormal functions and processes

In addition to representing relationships among process participants, our framework can represent the roles that participants have in a modeled system. We distinguish two types of roles: molecular-level functional roles (e.g., a role in translation) and roles in clinical disorders (e.g., the cause of cardiomyopathy). Each role is specified using a function/process code taken from the TAMBIS ontology. To represent dysfunctional molecular-level roles, we use an attribute, called role_present, which signifies whether the role is present, absent, or whether this information is unknown. For example, Figure 2 shows that three mutation of tRNA that exhibit the role of Misreading. The figure also shows tRNA mutations that have roles in the

cardiomyopathy disorder. Cardiomyopathy is one of the concepts from the clinical ontology, discussed later in this section.

Processes are represented using the Process Model component of our framework. We augmented the Workflow model with elements taken from the Object-Process-Methodology (OPM) [31], to create a graphical representation of the relationships between a process and the static components that participate in it, as shown in Figure 3. We used different connectors to connect a process to its input sources, output sources, and participants that do not serve as substrates or products (e.g., catalysts such as amino acid synthetase). We added a fourth type of connector that links a process to a chemical that inhibits the process (e.g., borrelidin). Figures 3 through 6 present details of the translation process and the processes leading to it. The figures show the normal process as well as processes that result in abnormal translation. We have considered only tRNA-related failures of translation. Detailed explanation of each Process Diagram is given in the legends. Figures 4 and 5 present the details of the translation process, depicted in Figure 3. Figure 4 represents the translation process according to the classical two-site model [5]. Figure 5 represents a recent model of the translation process [32]. The details of the process of tRNA binding to A-site, of Figure 5, are shown in Figure 6.

Insert Figures 3 - 6 here

The processes *Normal reading*, *Misreading*, *Frame shifting*, and *Halting*, shown in Figure 6, all have a process code of *Binding*, since in all of them, tRNA binds to ribosome that has occupied E and P sites.

The types of arrows that connect molecules to a process define their role as substrates, products, inhibitors, activators, or molecules that participate without changing their overall state in the framework (e.g., enzyme). The logical relationships among participants are specified in a formal expression language. For example, double-clicking on the *Misreading* process, shown in Figure 6, shows its participants, which are specified as:

((Shine dalgarno in E XOR tRNA0 in E) AND tRNA1 in P AND (tRNA2 that can bind to incorrect codon in Ternary complex XOR tRNA that has altered flexibility in Ternary complex) AND tRNA2 in A AND EF-tu AND GDP)

Representing high-level clinical phenotypes

Our clinical ontology relies on the UMLS, but does not include all of the concepts of the Metathesaurus. Instead, we are building our clinical ontology by importing concepts, as we need them. We add clinical concepts to the clinical ontology be creating them as subclasses of the semantic types defined by the Semantic Network. Each concept has a concept-name and a concept-code that come from the Metathesaurus, as well as synonyms. Figure 7 shows part of the clinical ontology. Figure 3 shows that mutated Leucine tRNA (in tRNA acceptor domain) and Mutated tRNA (in T domain) have roles in some forms of cardiomyopathy. Many tRNA-related diseases are also linked to mutations in protein components of mitochondrial respiratory chains. Proteomic studies in [28] provide a larger list of protein candidates. 20 identified proteins are shown to either overproduce (9) or be underrepresented (11) when mitochondrial genome has A8344G mutation (in $tRNA^{Lys}$), associated with Myoclonic Epilepsy and Ragged Red Fibers (MERFF) condition.

Insert Figure 7 here

Representing levels of evidence for modeled facts

Different facts that are represented in our framework are supported by varying degrees of evidence. It is important to allow users to know what support do different facts have, especially in cases of conflicting information. We therefore added a categorization of evidence according to the type of experimentation by which facts were established. The categorization includes broad categories, such as "in vivo", "in vitro", “in situ”, "in culture", "inferred from other species", and "speculative". Facts, such as the existence of a biomolecule, or its involvement in a process, are tagged with the evidence categories.

Querying the model

Using PAL we composed first-order logic queries that represent in tabular form relationships among processes and structural components. Table 1 shows a summary of all the query types that we composed. They are grouped into six categories that concern: (1) alleles, (2) functional roles and roles in disorder phenotypes, (3) reactions and their participants, (4) biological processes, (5) ability to reach a certain state of a modeled system, and (6) temporal/dynamic aspects of a modeled system. Queries that were especially interesting to us were (1) finding mutations that cause molecular-level processes and functions to be dysfunctional, (2) finding mutations that cause clinical disorders, and (3) finding processes that might be affected in a given disorder. Figure 7 shows the query and query results for the third query.

Insert Table 1 here

Simulating the model

As shown in Figures 4 and 5, we created two different models of the translation process: a historical model, and a current model. When we translated the workflow models into the corresponding Petri Nets, we were able to test predictions of these two models by showing that under certain concentrations of reactants, the different models resulted in different dynamic behavior, which produced different translation products. For example, when the mRNA contained a sequence of Asn-Leu-Asn (or in general, aa_1-aa_2-aa_1) and the system was initialized with a low concentration of Asn-tRNA, then protein translation proceeded in the classical two-site model but was halted in the current three-site model, which required Asn-tRNA and Leu-tRNA to be bound to the ribosome while a second Asn-tRNA bound the A-site. The Petri Net that corresponds to the workflow model of Figure 5 is shown in Figure 8. tRNA mutations were represented as colored tokens, belonging to the tRNA color set (See Figure 8), and mRNA molecules were represented as tokens belonging to the mRNA color set.

The Petri Nets derived from our workflow model can also be used for educational purposes. They can demonstrate: (1) concurrent execution of low-level processes within the translation process (e.g., tRNA molecules that were incorporated into synthesized proteins can be amino acylated and used again in the translation process), (2) introduction of mutations into synthesized proteins, and (3) the affect of certain dysfunctional components on pools of reactants (e.g., non-mutated tRNAs).

Insert Figure 8 here

4. Discussion

Deducing molecular mechanisms of disease based on molecular models is a very difficult problem. Even more complicated is the task of correlating genotypic variation to clinical phenotypes. A review by Florentz and Sissler [33] shows that, despite the accumulation of information about the positions of a large number of mutations within mitochondrial tRNAs, it is not possible to identify simple basic patterns for use in predicting the pathogenicity of new mutations. The multifaceted nature of effects produced by tRNA mutations is apparent from recent proteomics studies [29], and is emphasized in current reviews [34, 35]. The authors conclude that it is critical to examine not only the affected tRNA, but also its interactions, or relationships, with other compartmental components. These arguments emphasize the importance of a knowledge model able to integrate practical information at multiple levels of detail and from multiple experimental sources.

The knowledge framework presented here links genetic sequence, structure and local behavior to high-level biological processes (such as disease). The model provides a mechanism for integrating data from multiple sources. In our tRNA example, we integrated information from structural biology, genetics and genomics, molecular biology, proteomics, and clinical science. The information can be presented graphically as Process Diagrams or Participant/Role Diagrams. The frames that represent participants, roles, processes, and relationships among them contain citations to the original data sources.

Our model has several advantages, in addition to its ability to integrate data from different sources. First, we can define queries that create views of the model in a tabular format. The queries extract useful relationships among structures, sequences, roles, processes, and clinical phenotypes. Second, our model can be mapped in a straightforward manner to Petri Nets. We

developed software that automatically translates our biological process model into Petri Net formalisms and formats used by various Petri Net tools [36]. We have used available tools to qualitatively simulate a modeled system and to verify its boundedness and liveness and to answer a set of biological questions that we defined [36]. *Boundedness* assumes that there is no infinite accumulation of tokens in any system state. In our example, this corresponds to concentration of tRNA and mRNA molecules in a cell. *Liveness* ensures that all Petri Net transitions (which correspond to Workflow activities) can be traversed (enabled).

A disadvantage of our model is its need for manual data entry. Natural language processing techniques are not able to automatically parse scientific papers into the semantic structure of our ontology. The effort required to enter data into our model is considerable. The entry of a substantial set of data about all relevant cellular reactions and processes would require a major distributed effort by investigators trained in knowledge representation and biology.

5. Conclusion

One of the ultimate goals of proteomics and genomics engineering is to develop a model of the real cell, of its program responsible for different behaviors in various intra- and extra-cellular environments. Our long-term goal is to develop a robust knowledge framework that is detailed enough to represent the phenotypic effects of genomic mutations. The results presented here are a first step in which we demonstrate that the knowledge model developed in another context (malaria invasion biology) is capable of capturing a qualitative model of tRNA function. We have presented a graphical knowledge model for linking genetic sequence polymorphisms to their structural, functional, and dynamic/behavioral consequences, including disease phenotypes. We have shown that the resulting qualitative model can be queried (1) to represent the compositional properties of the molecular ensembles, (2) to represent the ways in which

abnormal processes can result from structural variants, and (3) to represent the molecular details associated with high level physiological and clinical phenomena. By translating the Workflow representation into Petri Nets we were able to verify boundedness and liveness. Using simulation tools, we showed that the Petri Nets derived from the historic and current views of the translation process yield different dynamic behavior.

Acknowledgements

The work was funded by the Burroughs-Wellcome Fund, and by NIH grants LM-05652 and LM-06422.

References

[1] M. Peleg, I. Yeh, and R. B. Altman, "Modeling biological processes using Workflow and Petri Net models," Bioinformatics, vol. 18, pp. 825-837, 2002.

[2] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," Int. Journal of Human-Computer Studies, vol. 43, 1995.

[3] S. Schulze-Kremer, "Ontologies for Molecular Biology," presented at Proceedings of the Third Pacific Symposium on Biocomputing, 1998.

[4] M. Ibba, C. Stathopoulos, and D. Soll, "Protein synthesis: twenty three amino acids and counting," Curr Biol., vol. 11, pp. R563-5, 2001.

[5] K. H. Nierhaus, "New aspects of the ribosomal elongation cycle," Mol Cell Biochem, vol. 61, pp. 63-81, 1984.

[6] D. N. Wilson, G. Blaha, S. R. Conell, P. V. Ivanov, H. Jenke, U. Stelzl, Y. Teraoka, and K. H. Nierhaus, "Protein Synthesis at Atomic Resolution: Mechanistics of Translation int he Light of Highly Resolved Structures for the Ribosome," Current Protein and Peptide Science, vol. 3, pp. 1-53, 2002.

[7] P. J. Farabaugh and G. R. Bjork, "How translational accuracy influences reading frame maintenance," EMBO Journal, vol. 18, pp. 1427-34, 1999.

[8] V. Volpetti, R. Gallerani, C. D. Benedetto, S. Liuni, F. Licciulli, and L. R. Ceci, "PLMltRNA, a database on the heterogenous genetic origin of mitochondrial tRNa genes and tRNAs in photosynthetic eukaryotes" Nucleic Acids Res., vol. 31, pp. 436-438, 2003.

[9] M. Peleg, I. S. Gabashvili, and R. B. Altman, "Qualitative models of molecular function: linking genetic polymorphisms of tRNA to their functional sequelae," Proc IEEE, vol. 90, pp. 1875-1886, 2002.

[10] L. Fisher, Workflow Handbook: Published in association with the Workflow Management Coalition, 2001.

[11] J. L. Peterson, Petri Net Theory and the Modeling of Systems. Englewood Cliffs, NJ: Prentice-Hall, 1981.

[12] W. M. P. v. d. Aalst, "The application of Petri Nets to Workflow Management," The Journal of Circuits, Systems and Computers, vol. 8, pp. 21-66, 1998.

[13] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, "An ontology for bioinformatics applications," Bioinformatics, vol. 15, pp. 510-520, 1999.

[14] C. Lindberg, "The Unified Medical Language System (UMLS) of the National Library of Medicine," J Am Med Rec Assoc, vol. 61, pp. 40-42, 1990.

[15] J. Odell, "Six different kinds of composition," Journal of Object-Oriented Programming, vol. 7, pp. 10-15, 1994.

[16] S. W. Tu and M. A. Musen, "Modeling Data and Knowledge in the EON Guideline Architecture," presented at Medinfo, London, 2001.

[17] H. M. W. Verbeek, T. Basten, and W. M. P. v. d. Aalst, "Diagnosing Workflow Processes using Woflan," The Computer Journal, vol. 44, pp. 246-279, 2001.

[18] D. CPN group at the University of Aarhus, "Design/CPN - Computer Tool for Coloured Petri Nets," 2002. http://www.daimi.au.dk/designCPN/

[19] M. Helm, H. Brule, D. Friede, R. Giege, D. Putz, and C. Florentz, "Search for characteristic structural features of mammalian mitochondrial tRNAs," RNA, vol. 6, pp. 1356-1379, 2000.

[20] M. Sprinzl and K.S.Vassilenko, "Compilation of tRNA sequences and sequences of tRNA genes," Nucleic Acids Res., vol. 33, pp. D135-D138, 2005.

[21] J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell, "The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs," BMC Bioinformatics, vol. 3, pp. 2, 2002.

[22] P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner, "SCOR: a structural classification of RNA database," Nucleic Acids Res., vol. 30, pp. 392-394, 2002.

[23] P. A. Limbach, P. F. Crain, and J. A. McCloskey, "Summary: the modified nucleosides of RNA," Nucleic Acids Res., vol. 22, pp. 2183-2196, 1994.

[24] "Online Mendelian Inheritance in Man, OMIM (TM)." McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000.. http://www.ncbi.nlm.nih.gov/omim/

[25] P. D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky[1], P. Kaipa, D. Ahrén[1], S. Tsoka[1], N. Darzentas, V. Kunin, and N. López-Bigas, "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes," Nucleic Acids Res., vol. 33, pp. 6083-6089, 2005.

[26] D. L. Wheeler, T. Barrett, D.A Benson., S.H Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J. U. Pontius, K.D. Pruitt, G. D. Schuler, L. M. Schrim, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the National Center for Biotechnology Information," Nucleic Acids Res., vol. 33, pp. D39-D45, 2005.

[27] M.C. Brandon, M. T., Lott, K.C. Nguyen, S. Spolim, S.B. Navathe, P. Baldi, and D. C. Wallace, "MITOMAP: a human mitochondrial genome database--2004 update," Nucleic Acids Res., vol. 33, pp. D611-613, 2005.

[28] W. T. Peng, M. D. Robinson, S. Mnaimneh, N. J. Krogan, G. Cagney, Q. Morris, A. P. Davierwala, J. Grigull, X. Yang, W. Zhang, N. Mitsakakis, O. W. Ryan, N. Datta, V. Jojic, C. Pal, V. Canadien, D. Richards, B. Beattie, L. F. Wu, S. J. Altschuler, S. Roweis, B. J. Frey, A. Emili, J. F. Greenblatt, and T. R. Hughes, "A panoramic view of yeast noncoding RNA processing," Cell, vol. 113, pp. 919-933, 2003.

[29] P. Tryoen-Toth, S. Richert, B. Sohm, M. Mine, C. Marsac, A. V. Dorsselaer, E. Leize, and C. Florentz, "Proteomic consequences of a human mitochondrial tRNA mutation beyond the frame of mitochondrial translation," J Biol Chem, vol. 278, pp. 24314-24323, 2003.

[30] V. Giudicelli and M. P. Lefranc, "Ontology for immunogenetics: the IMGT-ONTOLOGY," Bioinformatics, vol. 15, pp. 1047-54, 1099.

[31] D. Dori, "Object-Process Analysis: Maintaining the Balance Between System Structure and Behavior," Journal of Logic and Computation, vol. 5, pp. 227-249, 1995.

[32] S. Connell and K. Nierhaus, "Translational termination not yet at its end," Chembiochem, vol. 1, pp. 250-3, 2000.

[33] C. Florentz and M. Sissler, "Disease-related versus polymorphic mutations in human mitochondrial tRNAs: Where is the difference?," EMBO Reports, vol. 2, pp. 481-486, 2001.

[34] H. T. Jacobs, "Disorders of mitochondrial protein synthesis," Hum. Mol. Gen, vol. 12, pp. R293-R301, 2003.

[35] L. M. Wittenhagen and S. O. Kelley, "Impact of disease-related mitochondrial mutations on tRNA structure and function," Trends Biochem. Sci., vol. 28, pp. 605-11, 2003.

[36] M. Peleg, D.Rubin D, and R. B. Altman, "Using Petri Net Tools to Study Properties and Dynamics of Biological Systems". J Am Med Inform Assoc 2005;12(2):181-199.

Biographies

Mor Peleg is a Senior Lecturer at the Department of Management Information Systems at the University of Haifa since 2003. She received the PhD degree in Information Systems Engineering from the Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology in 1999. She received the MSc degree in molecular biology from the Faculty of Biology at the Technion in 1994. In 1999-2003 she has been a post-doctoral research fellow at Stanford Medical Informatics, Stanford University Medical Center. Her research interests include biomedical informatics, knowledge representation, clinical guidelines modeling, biological process modeling, and systems development methodologies. Dr. Peleg has developed the temporal version of the Object Process Methodology together with Prof. Dov Dori, and the Guideline Interchange Format, version 3, together with the InterMed Collaboratory. In 1990 she was awarded the Knesset (Israeli Parliament) excellent student award, in 1995 and 1996 she received the excellent teacher assistant award from the Technion, in 1997/8 she was awarded the Wolf Foundation prize for excellent doctoral students, and in 2005 she was awarded the American Medical Informatics Association (AMIA) New Investigator Award. Dr. Peleg is a member of the American Medical Informatics Association (AMIA).

Irene S. Gabashvili is a senior research scientist and the technical lead of computational bioscience program at Hewlett Packard Labs. She received her PhD in Biophysics in 1992 from the Institute of Physics, Georgian Academy of Sciences. She served as the head of a laboratory at the Center of Genetic Ecology, Georgia till 1995. She was a visiting scientist in the University of Quebec in Trois-Rivieres, Canada (1995), a postdoctoral fellow in the University of Texas Health Science Center, San Antonio, Texas (1995-1997) and a research scientist in the Laboratory of Computational Biology and Molecular Imaging, Wadsworth Center, New York State Department of Health (1997-2001). In 2001-2003, she was working in the area of bioinformatics in the laboratory of Dr. Russ Altman, Stanford University. She is a member of Biophysical Society, QSAR & Modeling Society, and the International Society for Computational Biology (ISCB).

Russ Biagio Altman is professor of genetics, bioengineering & medicine (and computer science by courtesy) at Stanford University. His primary research interests are in applying computing technology to basic molecular biological problems of relevance to medicine. He is developing techniques for collaborative scientific computation over the Internet, including novel user interfaces to biological data, particularly for pharmacogenomics. Other work focuses on the analysis of functional microenvironments within macromolecules and the application of nonlinear optimization algorithms for determining the structure and function of biological macromolecules, particularly the ribosome. Dr. Altman holds an M.D. from Stanford Medical School, a Ph.D. in medical information sciences from Stanford, and an A.B. from Harvard College. He received the U.S. Presidential Early Career Award for Scientists and Engineers, a National Science Foundation CAREER Award, and the Western Society of Clinical Investigation Annual Young Investigator Award. He is a fellow of the American College of Physicians and the American College of Medical Informatics. He is a past-president and founding board-member of ISCB, an organizer of the annual Pacific Symposium on Biocomputing, and an associate editor of the Bioinformatics journal. He directs the Stanford Center for Biomedical Computation, and won the Stanford Medical School graduate teaching award in 2000.

Table and Figure Legends

Table 1. Types of biological queries and motivating biological examples

Figure 1. Representing tertiary structure components. Normal tRNA is composed of five nucleic acid tertiary structure components. One of these components (tRNA Acceptor Domain) is shown in the middle frame. Each nucleic acid tertiary structure component is composed of segments of nucleic acid secondary structure components. The nucleic acid unpaired strand of the tRNA Acceptor Domain, which is a kind of nucleic acid secondary structure components, is shown on the right.

Figure 2. Part of the Participant/Role diagram showing molecules involved in translation and the roles that they fulfill. Individual molecules are shown as rectangles (e.g., tRNA). They are linked to domains (e.g., D-Domain) using dashed connectors. Biomolecular complexes are shown as hexagons (e.g., Ternary Complex), and linked to their component molecules using arrowhead connectors. Collections of molecules that share similar function or cellular location are shown as triangles (e.g., Elongation factors) and are linked to the participants that belong to them using connectors with round heads. Generalization relationships are shown as dotted lines (e.g., fMet-tRNA is-a tRNA). Functional roles are shown as ellipses that are linked to the participants that exhibit those roles. Clinical disorders that are associated with mutated participants are shown as diamonds (e.g., Cardiomyopathy) and are linked to the participants that exhibit roles in these disorders. The insert shows the details of the Misreading role. It is specified as Translation role (TAMBIS class) that is not present (role_present = false). Also shown are some of the participants that perform the misreading role.

Figure 3. A process diagram showing processes leading to translation. Ellipses represent activities. Ellipses with bold contours represent high-level processes, where ellipses without bold

contours represent low-level processes (that are not further expanded). The dark rounded rectangles represent routing activities for representing logical relationships among component activities of a process diagram. The router (checkpoint) labeled as ”XOR” represents a XOR split that signifies that the two processes that it connects to, are mutually exclusive. A XOR join connects the three processes shown in the middle of the diagram to the Translation process. Dotted arrows that link two activities to each other represent order relationships. Participants are shown as light rectangles. Arrows that point from a participant towards a process specify that the participant is a substrate. Arrows that point in the opposite direction specify products. Connectors that connect participants (e.g., amino acid synthetase) to processes and have a circle-head represent participation that does not change the state of the participant. Inhibitors (e.g., tobramycin) are linked to processes via a dashed connector. The details of the translation process are shown in figures 4 and 5.

Figure 4. A process diagram showing the details of the Translation process of Figure 4, according to the classical two-site model [5]. The symbols are as explained in the legend of Figure 3. After initiation, there is an aa-tRNA in the P-site (tRNA1 in P). During the “Binding to A-site and peptide bond formation” process a second aa-tRNA in ternary complex binds to the A-site. Two processes occur simultaneously at the next stage: movement of the second tRNA that bound to the A-site to the P-site, and at the same time, exit from the ribosome of the first tRNA that bound to the P-site. If the second tRNA, bound to the P-site, is of terminator type, termination occurs. Otherwise, the ribosome is ready to bind; the second tRNA to bind tRNA is now labeled as “tRNA1 in P” and another cycle of elongation can begin.

Figure 5. A process diagram showing the details of the Translation process of Figure 4, according to the model of Connell and Nierahus [32]. The details of the process "Binding of tRNA to A-site" are shown in Figure 6. After initiation, shine dalgarno is placed at the E site, and the first tRNA (tRNA1) is placed at the P site. Next, tRNA2 transiently binds to the A-site. This step is followed by three activities which are done concurrently: (1) exit from the E-site, of either shine dalgarno or tRNA0 bound to the E-site (at later stages of the elongation process), (2) binding to A-site followed by peptide bond formation, and (3) a routing activity (marked by an unlabeled round-corner square). The routing activity is needed for correspondence with the colored Petri Net that simulates the translation process, which needs to distinguish among the tRNA molecules that are bound to each of the three sites. At the next stage, tRNA2 at the A site shifts to the P site, and at the same time, tRNA1 at the P site shifts to the E site. If tRNA2 bound to the P-site is of terminator type, termination occurs. Otherwise, the ribosome is ready to bind; the second tRNA to bind is now labeled as "bound tRNA1", and the first tRNA to bind is labeled as "bound tRNA0", and another cycle of elongation can begin.

Figure 6. A process diagram showing normal and abnormal reading processes. Starting from a ribosome with tRNA in both E and P sites, four alternative processes can lead to ribosome with A-site occupied with tRNA: Normal reading, Misreading, Frame-shifting, and Halting. Symbols are as explained in the legend of Figure 3. The insert shows the process code (Binding) of the reading process, taken from the TAMBIS ontology.

Figure 7. A query that shows individual molecules that are involved both in disorders and dysfunctional processes. The results of this query may indicate which processes are involved in a given disorder. The query is shown on the left. The results are shown on the right. ?im1 ?process, ?role1, ?role2, and ?disorder, are frames that represent Individual Molecules, Processes, Roles, Roles, and Disorders, respectively. The query is written as a constraint. Instances that violate the constraint are returned. The predicate *(own-slot-not-null A B)* returns true if slot A of frame B is not null. The constraint looks for all individual molecules, which (1) have roles that are disorders, and (2) have roles that are dysfunctional processes or functions.

Figure 8. The Colored Petri Net that corresponds to Figure 5, showing the current three-site model of translation. Squares represent transitions, corresponding to Workflow processes. Ellipses represent places, corresponding to conditions that are true after a Workflow process has terminated. Text to the top-left of places indicates their allowed token type, which can be TRNA or MRNA. The values of tokens of TRNA type used in this figure are: Shine_Dalgarno, Initiator_tRNA, Terminator_tRNA, Terminator, Lys_Causing_Halting. Other token types that we use in our model, and are not shown in this figure, represent other mutations of tRNA molecules. The values of tokens of MRNA type are always “normal”. Text below places specifies initial placement of tokens in those places. Text above transitions indicates guarding conditions, which refer to token types. Text on connectors indicates token variables that flow on those connectors. The variables used are a, b, and c for TRNA tokens, and m for MRNA tokens. Transitions are also labeled t6..t15, in correspondence with query 5.2 of Table 1.