CHALKBOARD: ONTOLOGY-BASED PATHWAY MODELING AND QUALITATIVE INFERENCE OF DISEASE MECHANISMS

COOK, DL [1], WILEY, JC [2], & GENNARI, JH [3]

[1]*Physiology/Biophysics,* [2]*Comparative Medicine,* [3]*Biomedical & Health Informatics, University of Washington, Seattle, WA, 98195, USA*

We introduce Chalkboard, a prototype tool for representing and displaying cell-signaling pathway knowledge, for carrying out simple qualitative *reasoning* over these pathways, and for generating quantitative biosimulation code. The design of Chalkboard has been driven by the need to quickly model and visualize alternative hypotheses about uncertain pathway knowledge. Chalkboard allows the biologists to test *in silico* the implications of various hypotheses. To fulfill this need, chalkboard includes (1) a rich ontology of pathway entities and interactions, which is ultimately informed by the basic chemistry and physics among molecules, and (2) a form of qualitative reasoning that computes causal chains and feedback loops within the network of entities and reactions. We demonstrate Chalkboard's capabilities in the domain of APP proteolysis, a pathway that plays a key role in the pathogenesis of Alzheimer's disease. In this pathway (as is common), information is incomplete and parts of the pathways are conjectural, rather than experimentally verified. With Chalkboard, we can carry out in silico perturbation experiments and explore the consequences of different conjectural connections and relationships in the network. We believe that pathway reasoning capabilities and in silico experiments will become a critical component of the hypothesis generation phase of modern biological research.

1. Motivation

Molecular biologists must understand how biochemical reactions trigger downstream events leading to particular pathologies or phenotypes yet our signaling pathway knowledge is incomplete and volatile. Given a flood of high-throughput data, biologists increasingly depend on a myriad of well-organized, easily accessed data repositories [1-3] which, however, only provide the building blocks for generating and testing competing hypothetical pathway models of phenotypic expression. In this paper, we describe a candidate tool, Chalkboard, that allows biologists to easily build, revise and reason about pathway knowledge based on an ontological-based representation of the underlying chemistry and biophysics of pathway participants and reactions. Following Davis et al., we recognize that a knowledge representation is both the *declarative language* that captures knowledge (such as an ontology for pathway representation) but also a *inference method* that operates on the model [4]. The declarative language that allows one to state facts must be linked to the inference method that allows one answer questions about those facts.

Thus, Chalkboard is an hypothesis generation tool designed for ease-of-use that allows researchers to easily explore the behavior of hypothetical pathways in order to better direct *in vitro* or *in vivo* research. Chalkboard perturbation experiments graphically display the consequences of molecular activities and pathway links as a tool to identify downstream effects and inconsistencies with current knowledge. Furthermore, as quantitative pathway data become available, Chalkboard can automatically generate a set of quantitative differential equations in the JSim mathematical modeling language [5]. We demonstrate these capabilities by representing and testing a pathway model of amyloid precursor protein (APP) processing pathway are at the core of the "Amyloid hypothesis" of Alzheimer's Disease pathogenesis [6] as described in Section 4.

2. Overview of Chalkboard

Chalkboard is so named to emphasize its key features: First, one can create and modify models easily. Second, the system is designed for hypothesis generation and laboratory brainstorming—sharing, developing, and communicating hypothetical models with others. This contrasts with systems designed as repositories of consensus or authoritative models or datasets. Beyond physical a chalkboard, however, our Chalkboard models can be probed *in silico* to test ideas and predict outcomes as a guide to hypothesis generation.

2.1. Representing and visualizing biomolecules and their interactions

Chalkboard's representation of biomolecules, events and interactions are based on the BioD biological description language [7] which has evolved into an ontology organized around three major classes: *Entity*, *Action*, and *Functional attribute*. The *Entity* class represents basic cell biological entities such as compartments (*Compartment*; e.g., intracellular space, intranuclear space), molecules (*Molecule*, e.g., a protein or polynucleotide), and the functional domains of molecules (*Functional sites*, e.g., binding sites, catalytic sites). Chalkboard enforces rules for composing complex cell biological structures. For example, *Compartments* may be nested within *Compartments* but not within a *Molecule*; a *Functional site* can be a part of a *Molecule* but not *vice versa*. Chalkboard implements a number of primitive *Actions* to represent functional interactions between *Entities*. *Chemical flows* represent a variety of chemical processes such as *Bind reactions* (dimerization) and *Transporter flow* (across a *Compartment* boundary). *Actions* can be modulated (e.g., activated or inhibited) to represent the complex cell signaling logic. We also include "wildcard" classes (*Wildcard producer*, *Wildcard producer flow*, *Wildcard change action*) for representing entities and actions whose physical basis is unknown. *Functional attributes*

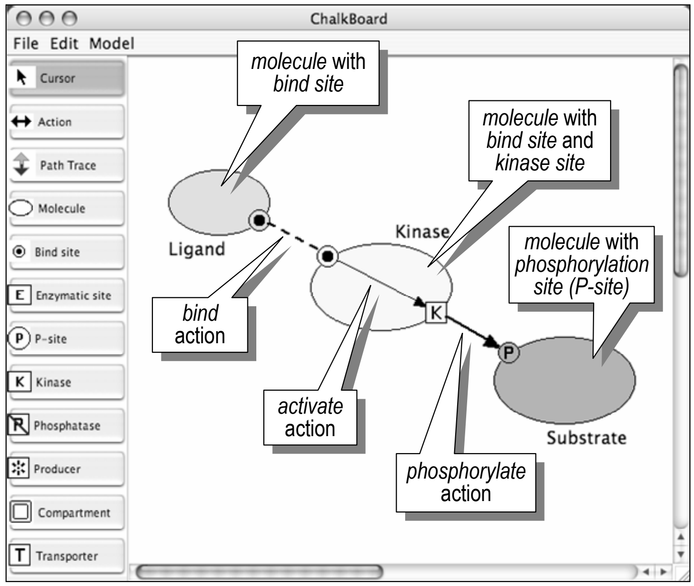

Figure 1. Annotated screenshot of the Chalkboard modeling environment showing the tool palette (left side) and a simple signaling cascade model.

represent the state attributes of *Entities* (e.g., *Concentration*) and *Actions* (e.g., *Rate* of a reaction) and provide the computational basis for qualitative reasoning (Chalkboard's inference method) and biosimulation code generation.

Our emphasis to date has been to create an ontology and computational system that (a) is based on formal views of anatomy and physiology [8, 9] , (b) is sufficient to carry out qualitative inference (see Section 2.2) and (c) can automatically generate mathematical biosimulation code as warranted by available data (see Section 2.3). As described in Section 4, our ontology will conform to standards for sharing pathway and biological knowledge [2, 10] while retaining its inference capability.

Figure 1 shows Chalkboard's graphical model editing environment that includes a model-drawing area and a tool palette for: a *Cursor*, a *PathTrace* tool (see Section 3.2), tools for installing *Entities,* and an *Action* tool for linking *Entities*. Model building is simplified because *Entities* and *Actions* are implemented as "smart" objects that enforce entity-composition and action-linking rules. For example, we built the simple signaling cascade in Figure 1 in steps: (1) Create and name 3 molecules with the "Molecule" tool. (2) Add to these molecules two *Binding sites*, a *Phosphorylation site* ("P-site") and a *Kinase site*. (3) Use the "Action" tool to install a *Bind action*, an *Activate action*, and a *Phos-*

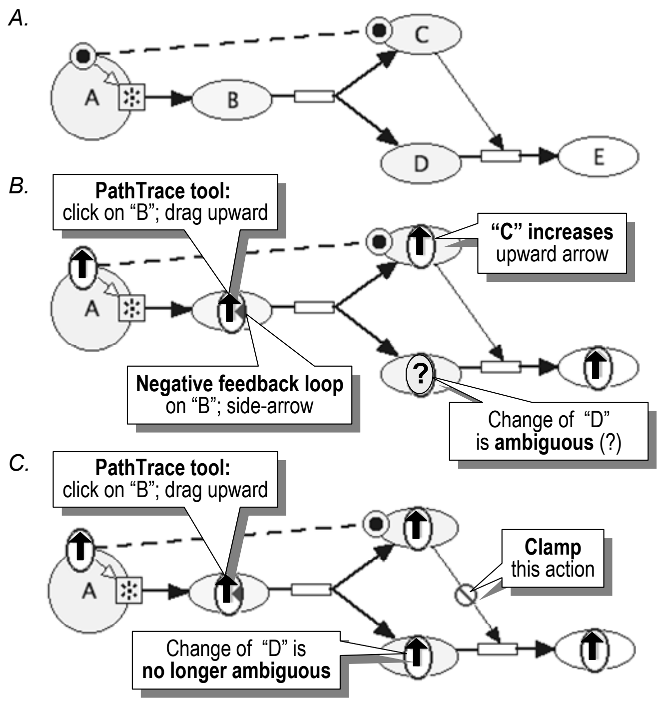

Figure 2. Chalkboard PathTracing applied to a simple metabolic model. **Panel A**. "A" produces "B" which dissociates into "C" and "D". D is transformed into "E" while C binds to a site on A that inhibits B production. The D-to-E transformation is activated by C via a Wildcard change action (solid-headed, single-weight arrow). **Panel B**. With PathTrace, the user clicks on B and drags ups to increment the amount of B. This increment propagates through the pathway and feeds back negatively on itself (a red side-arrow; positive feedback displays a green side-arrow). The change of D amount is ambiguous ("?" oval) because the increment of C due to the increment of B is counteracted by the activation of D transformation into E. **Panel C**. The Wildcard activation of the D-to-E reaction has been *clamped* (the slash sign; equivalent to a "functional knockout") so that the change in D amount is no longer ambiguous.

phorylate action. Chalkboard's context-sensitive linking automatically installs the correct *Action* for the *Entities* being linked. Even the limited set of primitive *Action* and *Entities* in Chalkboard's current ontology provides a rich and flexible vocabulary for creating models of complex cell biological systems. For example, activation actions can be either inhibitory (open arrowhead) or excitatory (solid arrowhead). In Section 4, we explore a biological example that includes proteolytic reactions.

2.2. Inference using PathTracing

Chalkboard's PathTrace tool allows researchers to carry out exploratory thought experiments in silico using an inference method that simulates qualitative responses to small perturbations of the system. Qualitative responses are displayed with 3 values (Figure 2): increase (upward arrow), decrease (downward arrow) or ambiguous ("?" oval). PathTracing also detects feedback loops as well as the effects of *in silico* "functional knockouts" by "clamping" an *Entity* or an *Action*.

PathTracing has three user-selectable modes: 1) Find all consequences of the perturbation of an index *Entity* or *Action* (as shown below). 2) Find only those feedback loops originating at an index *Entity* or *Action* ("Feedback only" mode; not shown). 3) Find only those pathways by which an index node affects any other preselected node (the "A-to-B" mode; not shown).

2.3. Architecture for PathTracing and biosimulation code generation

The computational architecture that underlies PathTracing also can be used to generate differential equation biosimulation code. Chalkboard *Entities* and *Actions* are endowed with *Functional attributes* (*FA*) that represent the values of their physical properties. For example, a *molecule* has a single *FA*, its *amount* (*amt*); how much of the molecule exists in the system. *Functional sites* have three properties: *amount* (assumed to equal to the *amount* of the site's parent molecule), *activity* (the fraction of *sites* in an active state), and *availability* (the amount of *active* sites). *Binding sites* are specialized with two additional *FA*s: *occupancy* (the fraction of *sites* occupied by ligand-) and *bound amount* (the *amount* of occupied sites).

As each *Entity* and *Action* is installed in a model, its corresponding *FAs* are created and linked via *Operators*, directed arcs that represent how each *FA* value depends (either directly or inversely) upon the values of other *FAs*. The resulting *Inference network* (e.g., Figure 3) is the basis for both PathTracing inference and for biosimulation code generation. PathTracing is accomplished by propagating tokens through the *Inference network* each delivering an incremental or decremental perturbation from one *FA* node to another. Incoming perturbations are stored and displayed as up- or down-arrows (Figure 2). A subsequent perturbation with a polarity opposite to a stored perturbation displays a "?" oval.

At *Inference network* bifurcations, tokens are cloned and launched into outgoing arcs (as at a *Bind site amt*). At network convergences (e.g., at a *Binding action's Jnet*), if an incoming perturbation replicates a prior perturbation then the incoming token is terminated because cloning it would simply replicate prior network traversals. To detect loops, tokens enlist an identifier for each traversed *FA* node so that if it detects itself it declares a feedback loop and terminates the

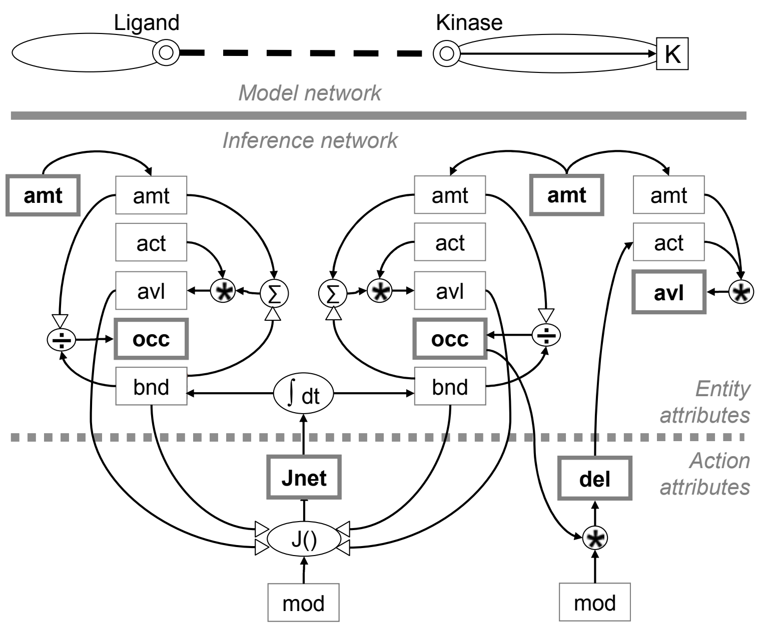

Figure 3. (Top) The ligand and kinase model from Figure 2. (Bottom) The *Inference network* (<u>not</u> visualized in the Chalkboard user-interface), *Functional attributes* (*FA*) for each *Entity* and *Action* are represented and linked by a network of *Operators* (white circles with mathematical symbols) and arcs (arrows) that represent the directed dependencies of attribute values on each other. PathTracing displays one "main" FA for each*Entity* or *Action* (bold frames). *FA's* in this model include:

amt = *Amount* of a *Molecule* or *Site*; molarity or concentration,
act = *Activity* of a Site; percent or fraction,
avl = *Available amount* = *act* x *amt*; molarity or concentration,
occ = *Bind-site occupancy* of a *Bind site*; percent or fraction,
bnd = *Bound amount* of a *Bind site*; molarity or concentration,
Jnet = Chemical *flow rate* of reaction; moles/s, concentration/s,
Del = *Change* of *Site* attribute; percent or fraction,
mod = *Modulator (of action)*; percent or fraction

token. Feedback loops are characterized as positive or negative according to the net polarity of perturbations in the token's list. Tokens are also terminated when they reach nodes with no outgoing arcs (as at each *occ* in Figure 3).

Chalkboard reuses the *Inference network* to automatically generate JSim [5] mathematical biosimulation code (not shown) that includes: (a) system state variables (one for each *FA* value) with default units, (b) algebraic or differential equations for each *Operator* (e.g., a rate equation), and (c) *Operator* equation parameters (e.g., reaction rate constants). The JSim system interprets Chalk-

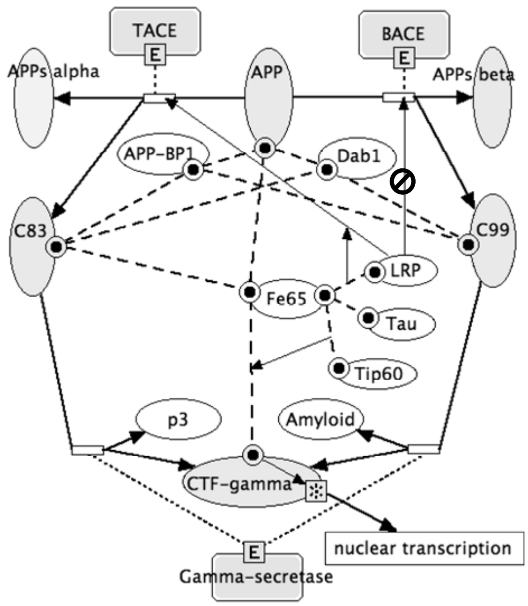

Figure 4. A view of APP proteolysis within Chalkboard where the action between LRP and the proteolysis by BACE is clamped. Under this condition, if more LRP is bound to Fe65, or if more LRP is available, then Amyloid production decreases.

board-generated code while parameter values are set by users at runtime.

3. A Chalkboard model of APP processing

Alzheimer's Disease is a pervasive neurodegenerative disorder associated with aging characterized by diffuse cortical plaques (neurofibrillary tangles) [6] whose primary constituent is a small peptide derived from the β-amyloid precursor protein (APP) [11]. The primary theory of Alzheimer's Disease etiology is the "amyloid hypothesis" by which elevated levels of β-amyloid production results in neuronal degeneration, cortical plaques, cognitive dementia, and ultimately death[6]. Effective therapy requires that scientists understand the complex events of APP proteolysis, in both normal and pathologic situations.

APP is a single-pass transmembrane protein that is sequentially proteolytically cleaved by enzymes to peptides (squares and ovals, respectively, in Figure 4). Primary cleavage occurs in the luminal/extracellular domain at the α-secretase cleavage site by metalloproteinases such as TACE [12] or at the β-secretase cleavage site by the atypical aspartyl protease BACE[13]. Subsequently, the remaining carboxy-terminal fragments of APP (C99 and C83 in

Figure 4) are cleaved by the heterotetrameric γ-secretase complex [14]. Cleavage of APP at the α- and γ-secretase sites (left hand side of Figure 4) liberates the APP extracellular domain (APPsα), p3 peptide, and the APP intracellular domain CTFγ (also called AICD)[15]. Alternatively, cleavage of APP at β- and γ-secretase sites, (right side, Figure 4) generates a soluble extracellular domain (APPsβ), an intracellular domain CTFγ, and amyloid β peptide[15]. CTFγ plays an important role in transcription. In particular, the heterotrimeric APP-CTFγ/Fe65/Tip60 complex functions as a nuclear targeted transcriptional regulator[16, 17].

It is unclear, however, how the CTFγ/Fe65/Tip60 complex affects neuronal survival[18, 19]. Furthermore, APP proteolysis by γ-secretase complex may be regulated by the APP-associated factor LRP[20] via Fe65[21] and also may involve the stimulation of either α-secretase or β-secretase cleavage[20, 22]. To test these possibilities, we have included the LRP/Fe65 binding in our Chalkboard model (Figure 4), and included LRP activation of *both* BACE and TACE proteolysis. Then, we clamped the effect of LRP on β-secretase cleavage (the slash sign), to show that the downstream effect is to decrease amyloid production.

The inherent complexity of the interactions among APP, the proteolytic processing enzymes, and the associated binding proteins is an arena in which a detailed modeling system such as Chalkboard would be extremely helpful. Potentially, Chalkboard could help provide valuable insights into predictions about both mechanisms of action and potential experimental manipulations that could guide the development of effective therapeutic approaches to treating AD.

4. Discussion and related work

Chalkboard is an ontology-based computational tool for representing biomolecular pathways using a graphical language and model editing environment to represent pathway models that can be analyzed qualitatively with a built-in PathTracing tool (Section 2.2) and analyzed quantitatively by exporting model simulation code (Section 2.3) to the JSim simulation system. As such, Chalkboard relates to several threads of computational research that deserve in-depth discussion beyond the scope of this paper. However, here we emphasize Chalkboards relation to three areas of pathway informatics research: Ontology research, qualitative inference, and quantitative analysis. We also address the tradeoffs between scalability and the rich biochemical representation we employ with Chalkboard.

4.1. Ontology-based representations of biomolecular pathways

The Chalkboard ontology continues to evolve from the BioD biological description language [7] concurrently with biomolecular pathway ontologies including BioPax [2], PATIKA [23], CellDesigner [24], and others. As expected there is considerable representational overlap that should, with community effort, be resolvable into a high-level ontology or, at least, alignment between related ontologies. We are committed to such efforts as advocated by others [10, 25].

We note, however, important representational differences, particularly in modeling molecular "states". Many ontologies consider different states of a physical entity (e.g., a molecule) to be separate entities (e.g., a molecule, its phosphorylated form, and its active form). Chalkboard takes an "object-oriented" view that a single entity *Molecule* can have *Functional sites* as parts and each part can have an independent operational state so that the state of a *Molecule* is specified by the values its own *Functional attributes* plus the *FAs* of its parts (e.g., *Occupied*, *Active*, etc.).

We adopt the *Functional attribute* approach because it maps well to both qualitative and quantitative analyses (Section 2.3). Furthermore, we suggest, the *Functional attribute* approach generalizes readily to other biophysical domains such as membrane biophysics (e.g., membrane potentials, conductances and currents), structural mechanics (e.g., elastance), and fluid flow (e.g., diffusive or bulk flows). We see this generalizability as a prerequisite for the integration of pathway knowledge and analysis into multiscale (molecules, cells, organs, organ systems, etc.), multidomain (biochemistry, biophysics, mechanics) models.

4.2. Qualitative inference and quantitative analysis

Qualitative reasoning tools in biological research have been driven by the scarcity and high cost of the quantitative datasets required for quantitative modeling. However, many representational schemes do not as yet, support qualitative inference (e.g., BioPax [2], CellDesigner [24]) and those that do use graph theoretic query methods (e.g., PATIKA [23]) and rule-based reasoning (e.g., BioCyc [3]) of state-based modeling. Chalkboard qualitative inference is more directly based on the principles of quantitative modeling by tracking the propagation of (small) perturbations through a network of essentially quantitative relationships.

The benefits of coupling graphical representations to the computational analysis of biological systems has long been recognized resulting in a variety of implementations including our own KineCyte [26] that integrates graphical modeling with biosimulation. Chalkboard, however, relies on existing simulation engines to interpret automatically-generated simulation code (currently, we use JSim but intend to support CellML[27] and SBML[28]). Although other molecular pathway representations (e.g., PATIKA, CellDesigner) may have sufficient rigor and expressiveness to export simulation code, to our knowledge, this has not yet available for existing simulation language[29].

4.3. Scalability and representational richness

We recognize trade-offs between Chalkboard's semantically-rich graphical view of biological pathways and less rich but more scaleable representations used by applications such as Cytoscape[30]. We believe that scientists need both sorts of tools—although Cytoscape is appropriate for coarse-grained visualization of large networks, only tools like Chalkboard, that use richer representations can capture notions of competitive binding, cooperative and anti-cooperative effects.

We recognize that Chalkboard will not be the only tool used by a researcher, and thus, we have the designed the system to export its models in a sharable format. Chalkboard models are saved in an XML text file that represents all model entities, model actions and their linkages in a form that can be read and parsed by other applications. Our plans more specifically include inter-operating with the BioPAX standard [2], (as much as possible, given the differences in modeling), as well as to CellML and SBML for simulation code.

5. Summary

We have argued that modern pathway researchers need tools for building and reasoning about causal models based on an *inference method*. Chalkboard is one prototype system that fills this need. The key characteristics of Chalkboard are: (1) The use of an expressive ontology of *Entities, Actions* and *Functional attributes* to model pathways at a based on the physics and biochemistry of inter- and intra-molecular interactions. And (2) Chalkboard's ability to carry out high-level symbolic qualitative inference (PathTracing) and to generate quantitative (JSim) simulation code allows users to avoids two pitfalls: (1) being tied to quantitative models whose utility and relevance are limited by the (typical) lack of quantitative data, and (2) over-simplified biochemical representations whose fidelity to actual biochemical processes is limited. We have introduced Chalkboard modeling environment and demonstrated its use analyzing a cell-signaling pathway with important scientific and clinical implications. The design of effective therapeutics requires a rigorous understanding of how modulation of a particular molecular entity would affect a distributed signaling system. As Chalkboard is designed to assess this issue, we suggest that use of Chalkboard modeling could facilitate the identification of appropriate pharmacogenetic therapeutic targets within Alzheimer's Disease and other human pathologies.

References

1. Joshi-Tope G. GM, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* 2005;33:D428–D432.
2. BioPAX – Biological Pathways Exchange Language Level 2. http://www.biopax.org. accesssed 2006
3. Karp PD, Ouzounis CA, Moore-Kochlacs C, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;33(19):6083-9.
4. Davis R, Shrobe H, Szolovitz P. What is a knowledge representation? *AI Magazine* 1993;Spring:17-33.
5. National Simulation Resource. http://nsr.bioeng.washington.edu/PLN. accesssed 2006
6. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 2002;297(5580):353-6.
7. Cook DL, Farley JF, Tapscott SJ. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol* 2001;2(4):RESEARCH0012.
8. Cook DL, Mejino JLV, Rosse C. The Foundational Model of Anatomy: a template for the symbolic representation of multi-scale physiological functions. *Medinfo* 2005;12.
9. Rosse C, Mejino JLV. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics* 2003;36:478-500.
10. Open Biomedical Ontologies. http://obo.sourceforge.net. accesssed
11. Glenner GG, Wong CW, Quaranta V, Eanes ED. The amyloid deposits in Alzheimer's disease: their nature and pathogenesis. *Appl Pathol* 1984;2(6):357-69.
12. Buxbaum JD, Liu KN, Luo Y, et al. Evidence that tumor necrosis factor alpha converting enzyme is involved in regulated alpha-secretase cleavage of the Alzheimer amyloid protein precursor. *J Biol Chem* 1998;273(43):27765-7.
13. Vassar R, Bennett BD, Babu-Khan S, et al. Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science* 1999;286(5440):735-41.
14. De Strooper B. Aph-1, Pen-2, and Nicastrin with Presenilin generate an active gamma-Secretase complex. *Neuron* 2003;38(1):9-12.
15. Selkoe DJ. Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* 2001;81(2):741-66.
16. Cao X, Sudhof TC. A transcriptionally [correction of transcriptively] active complex of APP with Fe65 and histone acetyltransferase Tip60. *Science* 2001;293(5527):115-20.

17. Baek SH, Ohgi KA, Rose DW, et al. Exchange of N-CoR corepressor and Tip60 coactivator complexes links gene expression by NF-kappaB and beta-amyloid precursor protein. *Cell* 2002;110(1):55-67.
18. Kinoshita A, Whelan CM, Berezovska O, Hyman BT. The gamma secretase-generated carboxyl-terminal domain of the amyloid precursor protein induces apoptosis via Tip60 in H4 cells. *J Biol Chem* 2002;277(32):28530-6.
19. Sastre M, Steiner H, Fuchs K, et al. Presenilin-dependent gamma-secretase processing of beta-amyloid precursor protein at a site corresponding to the S3 cleavage of Notch. *EMBO Rep* 2001;2(9):835-41.
20. Pietrzik CU, Busse T, Merriam DE, et al. The cytoplasmic domain of the LDL receptor-related protein regulates multiple steps in APP processing. *Embo J* 2002;21(21):5691-700.
21. Pietrzik CU, Yoon IS, Jaeger S, et al. FE65 constitutes the functional link between the low-density lipoprotein receptor-related protein and the amyloid precursor protein. *J Neurosci* 2004;24(17):4259-65.
22. Yoon IS, Pietrzik CU, Kang DE, Koo EH. Sequences from the low density lipoprotein receptor-related protein (LRP) cytoplasmic domain enhance amyloid beta protein production via the beta-secretase pathway without altering amyloid precursor protein/LRP nuclear signaling. *J Biol Chem* 2005;280(20):20140-7.
23. Demir E, Babur O, Dogrusoz U, et al. An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics* 2004;20(3):349-56.
24. Kitano H, Funahashi A, Matsuoka Y, Oda K. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* 2005;23(8):961-6.
25. Stromback L, Lambrix P. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics* 2005;21(24):4401-4407.
26. Cook DL, Gerber AN, Tapscott SJ. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc Natl Acad Sci U S A* 1998;95(26):15641-6.
27. CellML. http://www.cellml.org. accesssed 2005
28. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;19(4):524-31.
29. National Simulation Resource. http://nsr.bioeng.washington.edu/PLN. accesssed 2006
30. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498-504.