Semantic Web Applications in Neuromedicine: Problems and Prospects

Tim Clark
Massachusetts General Hospital
& Harvard Medical School

April 6, 2005
Courtesy of Tim Clark. Used with permission.
Neurodegenerative diseases impose huge social and personal costs

- Alzheimer Disease (5M people in US, 30M world)
- Parkinson’s (1.5M US, 6.3 M world)
- Huntington’s Disease (30K US, 150K US at risk)
- ALS (30K US)
- …others (CJD, etc)
AD: progressive, widespread loss of neurons …and consequently, brain function

Illustration: American Health Assistance Foundation

Courtesy of American Health Assistance Foundation. Used with permission.
Source: http://www.ahaf.org/alzdis/about/BrainAlzheimer.htm
Histopathology: NFTs and senile plaques

Image removed for copyright reasons.
Normal brain at autopsy


Courtesy of Alzheimer's Association. Used with permission.
AD brain at autopsy


Courtesy of Alzheimer's Association. Used with permission.
Alzheimer Disease research implicates diverse biological mechanisms

- genes (over 200 candidate genes published)
- environmental risk factors
- changes in cell function
- DNA damage
- misfolded proteins
- immune responses
- changes related to aging
- reduced regenerative capacity
- and others...
AD therapy development is highly interdisciplinary …

• etiological understanding
  – Genetics, genomics, proteomics, bioimaging, neuropathology, psychiatry, neurophysiology, cell biology…

• intersecting with
  – pharmacology, medicinal chemistry, animal studies and clinical trials
...and rather controversial

• A few current hypotheses on causation of AD
  – $\text{A}\beta$ protein direct toxicity to neurons
  – $\text{A}\beta$ protein indirect toxicity (“amyloid cascade”)
  – Defective cholesterol metabolism
  – Oxidative stress
  – Oxidative stress + abnormal mitotic signalling (“two-hit”)
  – Aluminum toxicity
  – Calcium signalling deficit
  – Disruption of white matter (oligodendrocytes and astrocytes)
  – NMDA receptor dysfunction

• “the truth is out there…” in the natural world
Can we build a useful knowledge base of research findings in AD?

• Classical knowledge bases resolve all internal contradiction
  – Tractable for reasoning from premises to conclusions
  – Composite of expert knowledge in a domain
  – Monotonic logic, “truth maintenance”

• What if the experts don’t agree?
  – The domain’s natural mode of reasoning is inductive
  – …and results of research are insufficient for rigorous proof
  – I.e., the domain is “underdetermined”
Core goals

- Contains useful current research results
- Can find “surprise” connections to other results
- Curation by true domain experts
- Computability & linkability of all statements
- Keep database current with the science
- Promote and incorporate active discussion
Technology

- Semantic Web (SW) technology is highly appropriate for this application
- This application is also an excellent demonstration platform for SW technology
Target: core pathways in AD

• Zero in on pathways of central relevance
• Proposed initial focus
  – genetics relevance (presenilin pathway)
  – therapeutics relevance (cholesterol pathway)
• Expansion of focus
  – multiple-disease relevance (protein misfolding)
  – Others as proposed by advisors
Curation by true domain experts

- We want real, leading researchers to curate
  - Experts must not become fulltime curators
  - Implies limitation of scope for each curator
- The toys in the cereal box (motivation)
  - Credit for ideas
  - Private whiteboard space in KB
  - Active collaboration space
- Ability to disagree & challenge statements
  - Part of the KB design
Web-deployed via AlzForum

• AlzForum is a global platform on AD
  – 70,000+ sessions per month
  – 32,000 visitors view 150,000 pages per month
  – Referenced by 7,700+ web sites
  – 2,000+ registered members
  – 3,000+ subscribers to newsletter

• The very top scientists in AD research serve each year on the AlzForum Editorial Board

• There is an active participant community
  – Online scientific dialogues and discussions
AlzForum scientific web community

Image removed for copyright reasons.
See http://www.alzforum.org/home.asp.
Semantic Web technology

• Vision outlined ~ 1998 (Berners-Lee)
• Next generation of WWW technology
• Major development milestones 2004
• can now begin to support large scale Semantic Integration of research results
• numerous life science applications
SWAN: Semantic Web Applications in Neuromedicine

• Pilot project in Alzheimer Disease
• Construct a broadly integrated KB of research results
• Curation by leading researchers
• Collaboration of researchers, computer scientists, industry and scientific web publishers
  – under auspices of W3C Semantic Web Activity
The SW-LS Technology stack (2005)

- LSID (Life Science Identifier)
- XML / XMLS
- RDF / RDFS
- OWL (Web Ontology Language)
Some challenges of Semantic Web application to neuromedicine

- Distributed identifier resolution
- Truth maintenance
- Computability
- Core ontological model
- Publication model / public-private ontology resolution
- Hypothesis representation
- Data provenance
- Socialization…
Distributed identifier resolution: LSID

- LSID - Life Science Identifiers (Clark, Martin & Liefeld 2004)
  - Globally resolvable, persistent, locally generated unique web identifiers
- Standardized by W3C & OMG
  - OMG standard issued 2004
  - Specialized W3C URN Namespace
- Increasing adoption in bioinformatics
  - Biopathways Consortium (LSID Authority Service)
  - Broad Institute (GenePattern)
  - UK eScience program (myGrid semantic grid project)
  - National Cancer Institute (caBIG cancer bioinformatics framework)
  - Genome Canada (bioMoby semantic service discovery project)
- Open source resolver software available
  - Lsid.sourceforge.net
Truth maintenance: Absolute or relative?

- Truth maintenance is typically conceived to be about *eliminating contradiction* in the KB
  - Okay on “boutique” scientific KBs
  - Does not scale as a process due to limitations of expertise
- Successful large KBs studiously avoid truth maintenance
  - Medline: correct bibliographic info = correct entry
  - Disregard truth or falsity of the science
  - “Let community process deal with it”, outside the KB
- Our approach: consciously *import contradiction* into the KB
  - “Relativization” (*reification*) of all statements
  - Private “idea incubation” sections of KB (the *Personal Whiteboard*)
  - Explicit statement publication to wide or narrow audience
...to a formal “computable” model

Dr. Joe Schmo

publish for release

believes @ 90%

Dr. Jane Sane

believes @ 100%

AlzForum

believes @ 60%

publish to recipient

ProteinA

activates

ProteinB

Statement1

Inactivates

ProteinC

Statement2

ProteinD ProteinE

Statement3

Complexes with

believes @ 65%

keep private

Dr. Betty Doe

publish for release

Dr. Mary Scary
Drs. Sane, Doe & Roe see these statements (public KB):
“Dr. Sane believes (100%) Protein C inactivates Protein A”
“Dr. Schmo believes (90%) Protein C inactivates Protein A”
Dr. Sane also sees these “private” statements:

“Dr. Sane believes (65%) that B complexes with D and E”
“Dr. Schmo believes (60%) that A activates B”
Core ontological model

• **[–constructive] models** (Hausser 2000)
  – Spekar-hearer is part of the model structure
    • goal is “to characterize truth”
    • previously seen as associated with science and mathematics

• **[+constructive] models** (Hausser 2000)
  – model-structure is part of the speaker-hearer
    • cognition seated in individuals and circumstances
    • previously seen as characteristic of language interpretation

• **explicit treatment of the hearer**
  – allows a collaboration network to be established
  – externalizes truth maintenance
Publication model and ontology alignment

**SWAN: Private and Public Concept Spaces**

*Transition from Private Hypothesis to Public Theory*

**PRIVATE knowledge**

**PUBLIC knowledge**

Time: $t_0$ to $t_1$
Hypotheses and ontology alignment

Formal Representation of Hypothesis in Metadata

PRIVATE knowledge
Exp(1)  Exp(2)  Exp(3)  Pub(p)

hypothesis

PUBLIC knowledge
Pub(1)  Pub(2)  Pub(k)

ontology (1)  ontology (2)  ontology (3)

identifier
about
description
notes
relationships
Data provenance

• Data level:
  – Where did this data come from?
  – Any known corrections/retractions/challenges?
• Information level:
  – What algorithms (and versions) were used?
• Knowledge level:
  – Who made this assertion? Any counter-assertions?
  – What publications /data are referenced?
• Constructive level:
  – Who may see /hear this assertion?
• Trust /security level:
  – Level of trust & security models for content & audience?
  – For the provenance software itself?
Socialization

• Changing the publication model for science…
• Data & publications = one’s career…
• KB curation and scalability…
• Trust model…
• Software development model…
To Sum Up

• Neurodegenerative diseases are huge problems
• Deep multi-disciplinary understanding required
• Technological basis for deep integration is here
• Potential huge benefit to researchers
• …and most importantly, patients

• We have to think very creatively about how we develop the next generation of Knowledge bases in neuromedicine
Collaborators

• Alzheimer Research Forum
  – June Kinoshita (co-PI), Elizabeth Wu

• MassGeneral Institute for Neurodegenerative Disease
  – Yong Gao, Georgios Asteris
  – Dora Kovacs, Lars Hernquist, Anne Young

• Brigham and Women’s Hospital
  – Dean Hartley

• MIT Computer Science and Artificial Intelligence Laboratory
  – Eric Miller, Ralph Swick

• IBM Advanced Technology Group
  – David Grossman, Sean Martin, Jordy Alboz

• HP Laboratories
  – Andy Seaborne, Steve Cayzer