Design of Integrated Translational Bioinformatics Systems

Barbara Mirel, DArs1; Benjamin J. Keller, PhD2,3; Mor Peleg, PhD3,4; Russ Altman, MD, PhD5

1 University of Michigan, Ann Arbor, MI; 2 Taudom Michigan Hospital, Toledo, MI; 3 University of Haifa, Haifa, Israel; 4 Stanford University, Stanford, CA

Abstract

Bioinformatics tools built for isolated tasks do not adequately support translational research; analytical workflows and integrated systems of tools are needed to support new interactive approaches to translational research. We examine user cognition and analysis patterns, their implications for system requirements, and biological knowledge requirements for causal analysis.

Introduction

• Bioinformatics tools are generally developed to perform an isolated task without full understanding of how a translational research workflow would use the tool in the context of a larger workflow. Efforts to build integrated systems of tools at NCBI make it necessary to understand how humans approach their research and how they try to employ these tools.

• Efforts to build integrated systems of tools at NCBI make it necessary to understand how users approach their research and how they try to employ these tools.

• Observations of users of integrated tools at NCBI are helping develop understanding of requirements for integrated systems of bioinformatics tools.

This effort meshes with ongoing work on biological modeling exemplified by the BioFirebox project.

User Observations and Interviews

• Methodology

Field observations and interviews over time Sample: 15 biomedical researchers

Interviews: Analysis: Software-supported problem-solving session aimed at formulating an hypothesis about disease mechanisms.

Data Collection: Think-alouds, task duration, actions, outcomes, impressions, goals, intentions, reasoning, judgments

• Data Analysis: Qualitative uncovering of patterns related to stages, tasks, behaviors, ways of knowing and reasoning

• Tool and Their Uses in Scientists’ Exploratory Analysis

Test: Michigan Molecular Interaction (MiMI) web interface and protein interaction database Used: Query on candidate genes or cancer, retrieve data about related protein interactions, literature, GO annotations, and literature-mined descriptions

Test: MiMI plugin in Cytscape.

Used: for interactively explore and manipulate networks of protein-protein interactions in Cytscape, complete with conceptual and quantitative data carried over from MiMI.

• Findings

• All scientists flowed through 4 stages, each with a dominant mode of reasoning:

  1. Confirming (classification)
  2. Afever Model Building (causal modeling/innovation)
  3. Separating the Wheat from Chaff (iterative validation)
  4. Building a Biological Story (exploration)

Classification

• All scientists flowed through 4 stages, each with a dominant mode of reasoning: Confirming, Afever Model Building, Separating the Wheat from Chaff, Building a Biological Story.

• Validation occurs throughout but in different forms in each stage.

• Classification reasoning is far better supported by bioinformatics tools than causal mental modeling and iterative validation.

• Novel insights into creative mechanisms of a disease depend on supporting users in integrating all modes of reasoning and cyclic stages of analysis.

• Support for novel insights requires improvements in tools, as follows:

  • Improve interactivity and control over manipulating data as needed for turning data into knowledge—i.e. based on observed patterns

  • Provide context, not just data read-outs, to guide inferences—e.g. emphasis, default groupings, “analysis in a keyword” that embeds domain knowledge

  • Provide surrogate inputs for evaluation acceptable to scientists when test statistics are not available or when biological knowledge is incomplete.

• Provide domain-based contextualizing information: Types of interactions, types of molecules, ability to import users' own data, overlay of protein interactions and pathways, overlays of protein interactions and disease-ome.

Matching Requirements to the User Model

• The user studies provide a basis for deriving a set of requirements based on the types of information inspected and tasks performed.

• Basic requirements relate to need to support:

  • All three kinds of reasoning:
    • Ability to verify outcomes from queries and previous stages
    • Ability to validate current information, and
    • Different approaches to moving among types of reasoning and verification/validation activities

• Within modes of reasoning requirements are more complex:

  • Example: Framing requirements for scientists’ different types of classification reasoning during the Separating the Wheat from Chaff stage

    • Scoping: Mechanisms for identifying:

      • Groupings by various forms of similarity—distinguishing between sharing of single versus multiple attributes

    • Patterns: temporal, localization, regulation, compound annotation, graph theoretic

      • Subsystems: internal structure relationships, normal vs abnormal behavior

      • Visualizations of multiple relationships across biological scale and attributes—e.g. a protein-protein interaction network painted by G2 Biological Process

• User interface for selecting, aggregating, and abstracting selections; filtering

• Display of information that allow the user to verify query results based on personal knowledge or literature presence

Supporting Modeling

• Many bioinformatics tools (including most NCBI tools) are focused on classification.

• The BioFirebox approach provides an insight into how modeling can be supported.

Example: BioFirebox Process Model of C-like Nucleosides

Model shows alternative processes involved in the cellular processing of C-like nucleosides, including the native C-nucleoside and the drug Gemcitabine – a C-analog (GFiC).

• Processive enzymes (green ovals) can transport the C-nucleosides into the cytoplasm, each using a different enabling transporter (blue rectangle): NENT1, NENT2, NENT3, and NENT7.

• C-like nucleosides (blue rectangle) is shown to be participating with a substance (link on all of the alternative processes, which converge on the SCORFDR rule.

• Two alternative processes follow: Deactivation of C-nucleosides or its phosphorylation.

• Phosphorylation (red arrow) is followed by alternative processes: Deoxycytidination, DeOXycytidination, or a second phosphorylation.

Structural Part of the Workflow

• Molecules that are part of the Gemcitabine/C-like nucleoside processing model are shown as blue squares. The rules that they play are shown as pink rules.

• The insert on the right shows the details of the Gemcitabine chemical. The cellular location is specified as “Nuclear”, – a term taken from the biomedical vocabulary that include TRANS and UM5.

Moving Forward

• Consideration of a cognitive view of user tasks leads to important insights in how to build integrated systems of tools.

• How possible to state more precise requirements for classification reasoning.

• Tools and scientific reasoning needs to be dynamically related, and an improved toolset will lead to cycle of identifying better targeted requirements followed by tool improvements.

• Few open source tools support contextualized causal reasoning for mental modeling and narratives.

• Well-targeted requirements will evolve in modeling and story-telling with user experience with prototypes based on the ideas such as the BioFirebox project.

• Remaining challenges include the design of visualizations for classification, and supporting the transition from classification to modeling.

Acknowledgements

This work was partially supported by National Institutes of Health: Grant # U54 DA021519.

References


