Information problems in molecular biology and bioinformatics

MacMullen, W. John and Denn, Sheila O. (2005). Information problems in molecular biology and bioinformatics. Journal of the American Society for Information Science & Technology 56(5), 447-456.

SILS Biomedical Informatics Journal Club

http://ils.unc.edu/bioinfo/

2005-09-06

Types of biomedical papers

- (literature) review - synthesis of prior work in one or more areas
- research / experimental - descriptive analysis, statistical analysis, hypothesis test
- 'project' / application - here's our new {tool, database, service} to address problem *x*
- theoretical - here's our theory about phenomena *x*, based on *y*
- editorial / position paper – makes recommendations, advocates
- ('book') review (JMLA, e-streams, etc.)

A Conceptual Map of Information Science Applications to Molecular Biology

Sheila O. Denn[1] and W. John MacMullen[1,2]

[1]School of Information and Library Science; [2]School of Medicine, Program in Bioinformatics and Computational Biology
The University of North Carolina at Chapel Hill

CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360 Email: {denns,macmw}@ils.unc.edu

Overview

While bioinformatics has generated a great deal of study within both the molecular biology and computer science research communities, there has been relatively little research on bioinformatics within information and library science, despite the fact that there are a number of clear opportunities for information scientists to make significant contributions. We believe that this is in part a direct result of uncertainty as to what bioinformatics actually is, and the nature of the problems in molecular biology that could be addressed.

This presentation seeks to reduce the ambiguity of bioinformatics and to enumerate classes of problems in molecular biology that can be addressed by information scientists. We provide a conceptual mapping of information science research areas to a generalized model of an experimental cycle in molecular biology, providing granularity of sub-tasks and citations to specific application examples, to illustrate potential "insertion points" for ILS researchers.

Methodology

We manually reviewed approximately 130 library and information science journals to find papers related to ILS research in the life sciences. We also reviewed the leading bioinformatics journals, and molecular biology journals that frequently publish bioinformatics-oriented articles. We searched PubMed MEDLINE for articles with the keywords or index terms "(computational biology or bioinformatics) and (information science)".

Results

We selected a subset of articles that we believe represent a broad spectrum of opportunity for ILS researchers. We mapped the articles to a model of a molecular biology experimental cycle, and to basic and applied ILS research areas. The applied research areas are applicable to multiple steps in the experimental cycle, and the basic research areas permeate the discipline. The distribution of citations reflects in part the major problems of bioinformatics: the integration of databases, the standardization of terminologies; and the problems of searching for, retrieving, synthesizing, summarizing, and visualizing information.

Discussion

While work is being done in the above areas, they are by no means solved problems. The nomenclature problem in molecular biology worsens in proportion to each new non-standard name assigned to a gene or gene product, just as the data integration problem worsens with each new data set that becomes available without standardized metadata frameworks.

Other gaps where little research has been done to date include:

- User task- / goal- / or problem- analysis: In general, more research is required to understand what types of problems investigators are trying to solve, and what tasks they are usng to do so.
- HCI: What types of interfaces would be most beneficial for these tasks?
- Communication: What would facilitate better communication and collaboration among researchers?
- Literature: What can be done to address the worsening problems of fragmentation of the literature and proliferation of vocabularies?

Citations are also largely clustered around experiment cycle elements that are post-data collection, suggesting that more work could be done on earlier steps, such as systems analysis and design during experimental design. While these are all challenges for biology, they are also opportunities for ILS to apply its expertise to a domain with interesting and challenging problems.

Basic ILS Research Areas

Can potentially encompass all parts of the experimental process

Domain Analysis
Domain analysis in this context is the study of particular subject disciplines or fields as communities of discourse, with the goal of using such analysis to inform the design of information systems and services targeted to users within such a discipline. Hjørland (2002) counts 11 major research approaches within this area, including subject gateway creation, classification, indexing and retrieval, user studies, and bibliometrics, among others.

Cotter, et al., 2000 El Kader, et al., 1998 Gilbert, 1991 Hjørland, 2002 Hjørland & Albrechtsen, 1995 Kitano, 2002 Lascar & Mendelsohn, 2001 Lenoir, 1999 Lynch, 1999 Norman, 1999 Ouzounis, 2002 Palmer, 1999a Pierce, 1999 Poinçot, 2000 Siegfried, 2000 White & McCain, 1998

Information Seeking & Use
This area has gone under a number of different names with subtle differences in meaning, but as Todd (1999) asserts, in general this cluster of concepts is concerned with "people and information coming together; it is about people 'doing something' with information that they have sought and gathered themselves or provided by someone else." Methodologies in this area have included intensive studies of user information needs, how these needs are expressed, how users interact with finding aids of various sorts, and what sorts of tasks they are trying to accomplish using information within their subject domain.

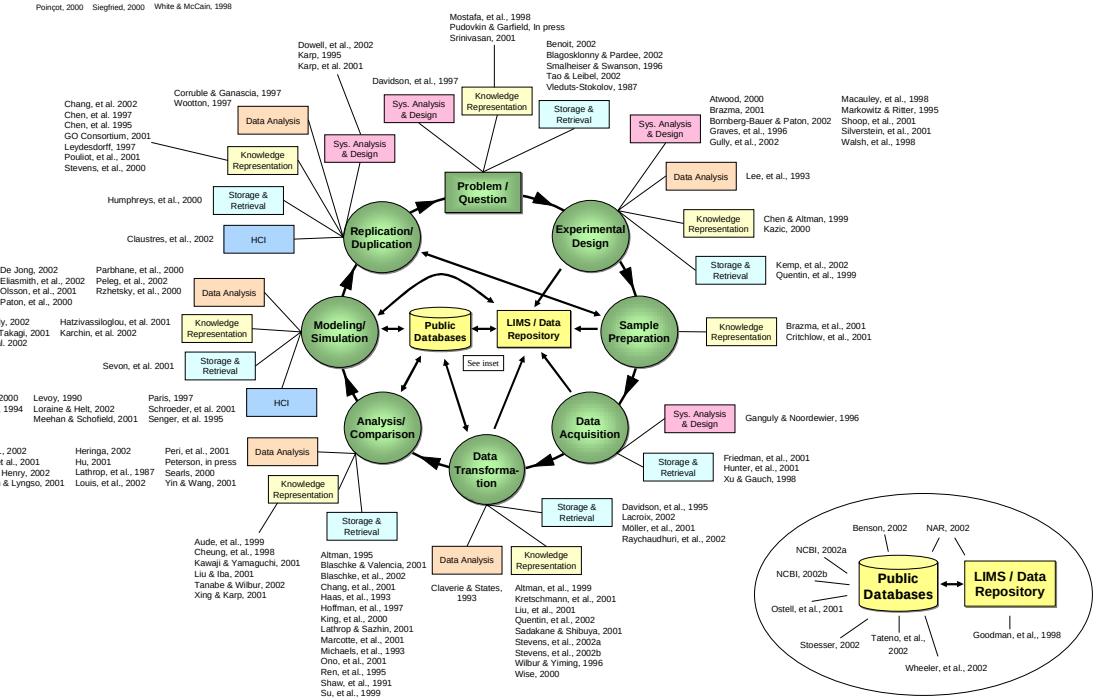
Arte, 2001 Bowden & DiBenedetto, 2002 Cole & Bawden, 1996 Hurd, et al., 1999 Newby, 2000 Palmer, 1999b Sinn, 2001 Stevens, et al., 2001 Todd, 1999 Yarfitz & Ketchell, 2000

Communication
This area is concerned with how information is communicated, especially through technological means. Research areas within communication include generalized models of communication (such as that of Shannon and Weaver, 1949), models of communication roles played by members of groups or organizations, models of the impact of communication on the adoption of technology, and the impact of information technology on communication and learning, including computer-supported cooperative work.

Sakai, 2001 Shannon & Weaver, 1949

Theories of Information
The study of information in its various forms and contexts is perhaps the most abstract component of the ILS research portfolio. Research here investigates information as thing, as process, as communication, and other approaches (Shannon & Weaver, 1949; Shannon, 1940; Losee, 1990). There are interesting possibilities to explore in molecular biology, such as the emergence of diversity from a small number of initial components and states, and the role of information in storage, transmission, and error correction in the genetic code.

Losee, 1990 Shannon, 1940 Shannon & Weaver, 1949 Stepanyants, et al., 2002



Applied ILS Research Areas

Are applicable at particular points in the experimental process

Systems Analysis & Design
This area is concerned with the efficient and effective modeling of systems, processes, information needs, user requirements, and information flows. Other components include database design and data modeling, system implementation, and the evaluation and modification of systems. Since most molecular biology work is based in small labs with little automation, significant analysis and design work is required when labs need integrated laboratory information management systems (LIMS) to handle high-throughput workflows.

Data Analysis
This category contains activities that involve manipulating content, such as modeling (including simulation and mapping; e.g., protein folding, and gene mapping), prediction (e.g., gene prediction and protein structure prediction), and comparison (e.g., sequence alignment). Since much of information science has historically not focused on content or meaning, this is an area where there are many opportunities to explore the application of methods to advance knowledge within a specific domain.

Knowledge Representation
This category contains a wide range of activities that are generally concerned with the content, semantics, and organization of information and its management, including classification, indexing, metadata, controlled vocabularies (e.g., thesauri and ontologies), and the annotation or curation of collections. The development of standardized and extensible knowledge representation frameworks will be absolutely critical to integrating the "islands of data" that exist in molecular biology today.

Storage & Retrieval
Storage and retrieval is considered to be one of the core areas of information science research. Storage and retrieval research is concerned with the issues of how data is stored within systems such that it can be efficiently retrieved at a later point. Areas of research within storage and retrieval include the design of data structures for the purpose of retrieval, algorithms for retrieving information based on user input, query languages and optimization, data mining, and the integration of retrieval results from across different storage systems.

Human-Computer Interaction
HCI is concerned with how information is presented to human users and what mechanisms are given to users to manipulate that information. Research areas within HCI include user interface design, visualization of data, and information presentation across different electronic devices. The concern here is not so much on the organization of the data, but how to transform the data so that it can best be understood and manipulated by human users, so there are aspects of cognitive psychology, human visual processing, and ergonomics that inform this research area.

Molecular Biology Experimental Cycle

Problem / Question
Much of molecular biology is question- or problem-driven, with experimental results and analysis feeding back and refining questions. One role for ILS is to understand the questions researchers are interested in answering, and ensure that researchers understand the ways in which ILS methods can facilitate investigation.

Experimental Design
The complexity and expense of molecular biology experiments requires precise and complete design. Research opportunities for ILS here include the design of open, integrated LIMS systems, the determination of what metadata elements to be captured during the experimental cycle, and models for the different types of data involved.

Sample Preparation
Some sample preparation procedures are highly complex, multi-step and multi-instrument processes that involve large amounts of metadata. This category encompasses both traditional "wet lab" and data-dependent "in silico" experiments.

Data Acquisition
Data are frequently acquired via automated instruments whose output is stored in a file or database as time series or other multivariate, high-dimensionality data. Many instruments feed data directly to the laboratory information management system (LIMS).

Data Transformation
Frequently, the raw data from instruments must be transformed prior to analysis; e.g., in gene expression experiments with microarrays, image analysis is required to translate the intensity of fluorescent tags into numeric values. This is often a multi-step process that involves converting the data from one format in one database to a different format in another.

Analysis/ Comparison
Analysis and comparison of experimental data increasingly involves the acquisition of known data from public databases (see inset) for comparative purposes (e.g., to infer structure or function). Key roles for ILS are the development and integration of metadata standards and the extraction of contextual data from public databases for annotation.

Modeling/ Simulation
This step makes use of computational tools to create representations of biological structures. As with Analysis / Comparison, this often involves retrieval of data from public databases. This is an important area for HCI research and development.

Replication/ Duplication
To test validity and accuracy, experiments are often repeated by the same or different investigators. Possible ILS roles include creating infrastructures that facilitate data normalization, synonym resolution, and database integration, as well as standardized test corpora to evaluate the performance of analysis tools.

This poster and the full reference list are available at: http://ils.unc.edu/~macmw/asist

Key points

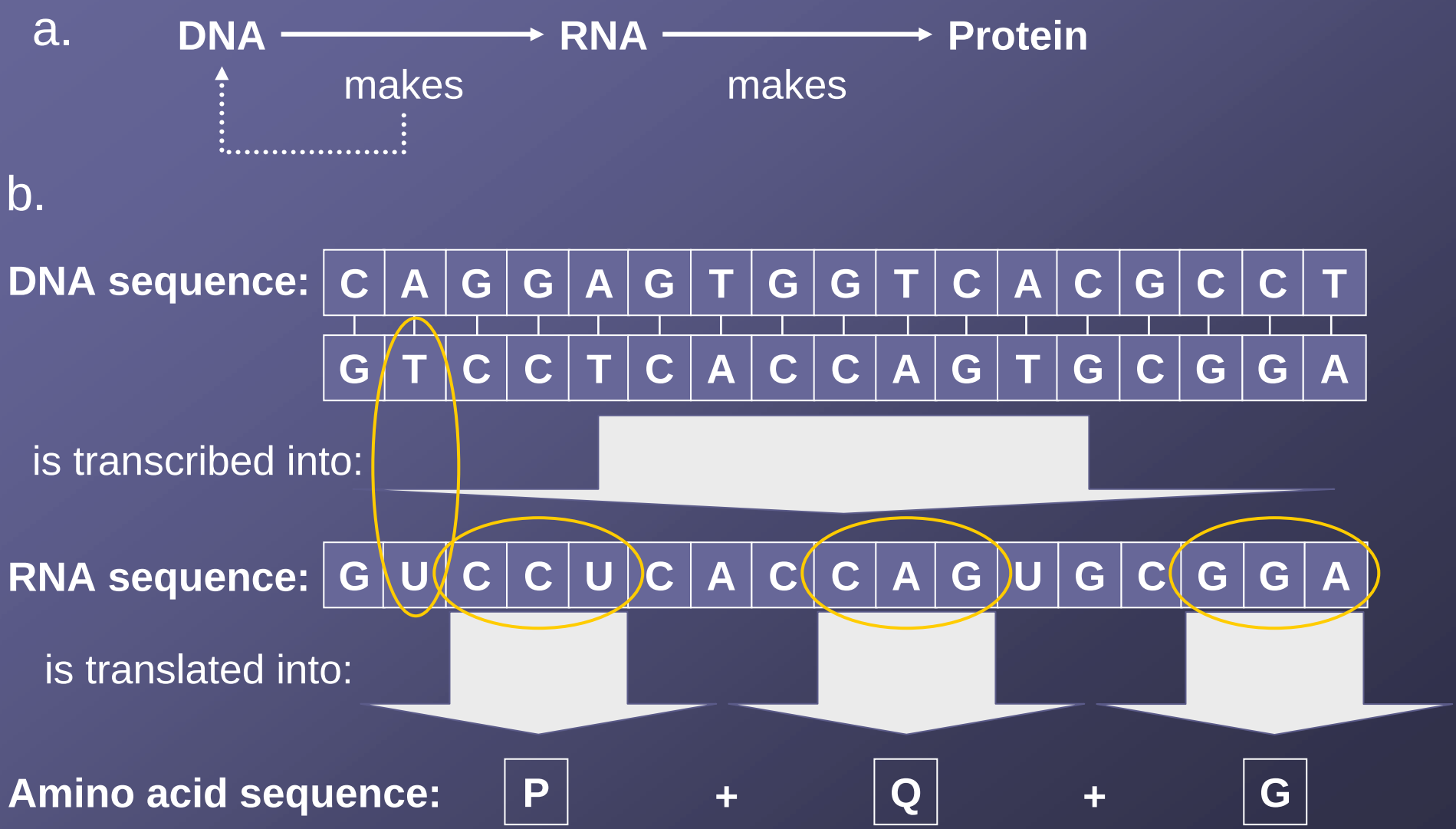
- Historically, it has been very difficult to acquire data in molecular biology
- Biomedical research has largely been driven by technological advances, which have led to greater physical and conceptual resolution
- Now that data exists in large quantities, research can move from reductionist, descriptive analysis to synthetic / integrative approaches to understanding functions, processes, and relationships

Information metaphors

- “From an information perspective,the general goals of molecular biology are to understand how the generation, communication, and interaction of biological information results in the creation and ongoing operation of living organisms [448].”
- Storage, retrieval, editing, transcription, translation, recombination, frame-shifting, noise, (non-)coding, error detection, messages, networks…

Two views of the Central Dogma

Fig 1, p. 449



Classes of Problems in Molecular Biology

- Structure
- Function
- Communication

Tasks in Molecular Biology

- Sequence Alignment
- Structure Prediction
- Function Prediction
- Comparative Genomics, Proteomics, and Metabolomics

ILS Research Streams and Insertion Points

- Information-Theoretic Approaches to Modeling and Measuring Biological Information
- Information Needs and Information-Seeking Behavior of Molecular Biologists
- Knowledge Representation Issues in Molecular Biology
- Data-Literature Integration
- Data-Literature Mining/Discovery Support Systems
- Visualization Tools and Interface Design Problems
- Library and Information Services to Support Molecular Biologists