Bioinformatics Panel Presentation

Peter D. Karp, Ph.D.

Director, Bioinformatics Research Group

SRI International

Menlo Park, CA

pkarp@ai.sri.com

My Background

- **Direct bioinformatics research group of 6 people in Artificial Intelligence Center at SRI International**

- **Stanford Computer Science Ph.D., 1989**
- **1.5 year post-doc at National Institutes of Health**
- **At SRI from 1991-present**
- **Vice President at DoubleTwist Inc. from 1997-1999**
- **Consulting assistant professor of Medicine at Stanford, 1994-present**

Introduction

- **What are**
  - Bioinformatics
  - Computational Biology
  - Biomedical computing

- **(Computer-Science * j) + (Biology * k)**
- **j = 1 - k**
- **0 > j < 1**

- **Education, research results, journals, funding sources, conferences, collaborators**

$j = .9 \; ; \; k = .1 \; ;$

- **Computer scientist who**
  - Performs computer science research in the context of biological problems
  - Designs computational paradigms based on biological systems
  - Earned Ph.D. in Computer science
  - Publishes in computer science journals only
  - Funded by NSF Computer Science

  - May or may not ever actually solve a biological problem
  - May or may not have biologist collaborators

$j = .1 \; ; \; k = .9 \; ;$

- **Biologist who**
  - Applies existing bioinformatics software to solve biological problems
  - Earned Ph.D. in biological sciences
  - Programs in Perl and SQL
  - Publishes in biology journals
  - Funded by NSF Biology, NIH, DOE
  - Might have taken a few computer science classes
  - May have developed some programming proficiency in other languages
- **My terminology: Computational biologist, not bioinformatics researcher**

$j = .3 \; ; \; k = .7 \; ;$

- **Interdisciplinary researcher who**
  - Develops a biological database or its supporting software, develops software for genome analysis or visualization
  - Develops sophisticated software to solve challenging biological problems
  - Earned Ph.D. in biological sciences, M.S. in computer science
  - Publishes in a mix of bioinformatics and biology journals
  - Funded by NIH, NSF, DOE

$j = .7 \; ; \; k = .3 \; ;$

- **Interdisciplinary researcher who**
  - Develops novel bioinformatics algorithms, ontologies
  - Uses state of the art computer science, or performs computer science research, to solve biological problems
  - Earned Ph.D. in computer science, B.A. in biology
  - Publishes in a mix of bioinformatics, computer science, and biology journals
  - Collaborates with biologists
  - Funded by NSF, NIH, DOE
  - Can’t find a job

Typical Mistakes Made by Computer Scientists New to Bioinformatics

- **Develop a beautifully engineered program that uses an elegant algorithm to rapidly solve the wrong problem**

- **Underestimate the importance of content**
- **Discovery = Algorithms + Databases**

How Changes in Computer Science Education Can Help Bioinformatics

- **Most natural scientists have little understanding of computer science**
  - Computer science is programming
  - Cannot appreciate the value that computer scientists bring to bioinformatics
  - Complexity of software engineering

Database Education

- **Science in the 21st Century is information intensive**
- **Over 300 databases in bioinformatics**
- **The database area of bioinformatics is where practice falls farthest behind the state of the art**
  - Few bioinformaticians trained in databases, knowledge representation, ontologies, formal languages
  - Little use of commercial DB technology until recently
  - Considerable design flaws in many DBs
  - Elementary mistakes made over and over
  - Dependency of databases on history
  - Database expertise vs mathematics expertise
- **All natural scientists should be educated in the area that spans databases, AI knowledge representation and ontologies, formal grammars, data models**

My Research

- Symbolic systems biology
- Encode biological theories in declarative form
- Knowledge base describing *E. coli* genome, proteome, metabolic pathways
- Algorithms and ontologies for metabolic pathways
- Algorithm for predicting the metabolic pathways of an organism from its genome
- Research in integrating knowledge bases and databases