Educating problem-solving skills for Bioinformatics research

Haixu Tang

School of Informatics, Indiana University
What skills are we looking for?

Technical Skills:
- Languages:  Java, SQL, PL/SQL, C,C++, HTML, PHP, JavaScript, JSP, VB(.Net), XML.
- Tools:  Eclipse, Toad, MS-Office softwares, MS Visual Studio, Dreamweaver.
- Servers:  Weblogic, WebSphere, JBoss

Technical Skills:
- Computer Languages & Technologies:  Core Java, J2EE, Swings, Struts, Servlets, EJBs, JSP, PL/SQL, XML, IBM Portlets, JSR 168 Portlets, Hibernate, Collections Framework, JDBC, RMI, Tiles, AJAX, UML, Teamsite, LDAP, Siteminder.
  - Application Servers:  WebSphere Application Server (WAS), ColdFusion, WebSphere Portal Server (WPS5.1), Oracle9ias Application Server and BEA Weblogic Application Server.
  - Databases:  Oracle8i & 9i, Microsoft Access.
  - Web Designing & Tools:  HTML, DHTML (Dynamic), JavaScript, Extensible Stylesheet Language Family (XSL), Cascading Style Sheets (CSS), Document Object Model (DOM), Macromedia Dreamweaver.
  - Job Schedulers:  Appworx 5.1 & Autosys.
  - Load Runner:  Mercury Load Runner, JUnit.

Software Tools:  WebSphere Studio Application Developer (WSAD), Rational Application Developer (RADv6), TOAD8.6 (for Oracle), Lotus Notes, SSH Client, Remedy, Rational Rose, Edit Plus 2, Textpad & JAD Decomplier

SECANT Workshop, Purdue University, 10/30/2008
What is missing?
technique skills vs. problem solving skills

• Information technology
  – Programming-driven
  – Many inexperienced users
  – Robust, user-friendly, scalable, modular
  – General models
  – Challenge: engineering

• Scientific computing
  – Problem-driven
  – A few experienced users
  – Accurate, efficient, specific (often novel) models
  – Challenge: problem solving

Domain knowledge is not a hurdle (at least in bioinformatics): teaching a computer scientist biology is usually easier than teaching a biologist computer science.

SECANT Workshop, Purdue University, 10/30/2008
Genome science: a revolution in biology

- Classical Biology
  - Hypothesis
  - Data
  - Knowledge
  - Hypothesis driven approach

- Genome Science
  - Data
  - Hypothesis
  - Knowledge
  - Data driven approach

SECANT Workshop, Purdue University, 10/30/2008
Bioinformatics: in the driving seat

- Classical Biology
- Data analysis

Hypothesis
↓
Data
↓
Knowledge

Hypothesis driven approach

- Genome Science
- Data mining

Data
↓
Hypothesis
↓
Knowledge

Data driven approach

SECANT Workshop, Purdue University, 10/30/2008
Problem solving skills for bioinformatics research

• Key: computational thinking about the data
  – High dimension & large amount
  – Objective: generating hypothesis

• Examples
  – Data visualization
  – Simulation
  – Data mining: rule discovery, classification, clustering, etc
  – Statistics: hypothesis testing, etc
  – Modeling: probabilistic modeling, etc
Data visualization

• What plot?
  – Scatter plot, bar plot, distribution, heat map

• Data representation
  – Vector, binning (density), ratio (log ratio)

• High dimension data

• Inhomogeneous data
  – Data integration

• Applying domain knowledge
Computational thinking vs. quantitative (mathematical) thinking

• Data-centered vs. hypothesis centered
• Statistics
  – Hypothesis tests
    • Model-based
    • Parametric vs. non-parametric
  – Permutation tests: simulation
• Modeling
  – Theoretical (i.e. mathematical) models
    • analytic solutions
  – Simulated models
    • Numerical solutions
  – Probabilistic models
    • NN, HMM, BN, etc.
Genome science: key advancements

• High throughput biotechnologies
  – Genome sequencing techniques
  – DNA microarray
  – Mass spectrometry
• Large-scale experiments
  – HGP, HapMap
  – Omics / Systems Biology
• Massive data generation, storage, exchange and analysis
  – Bioinformatics