Current Challenges in Bioinformatics

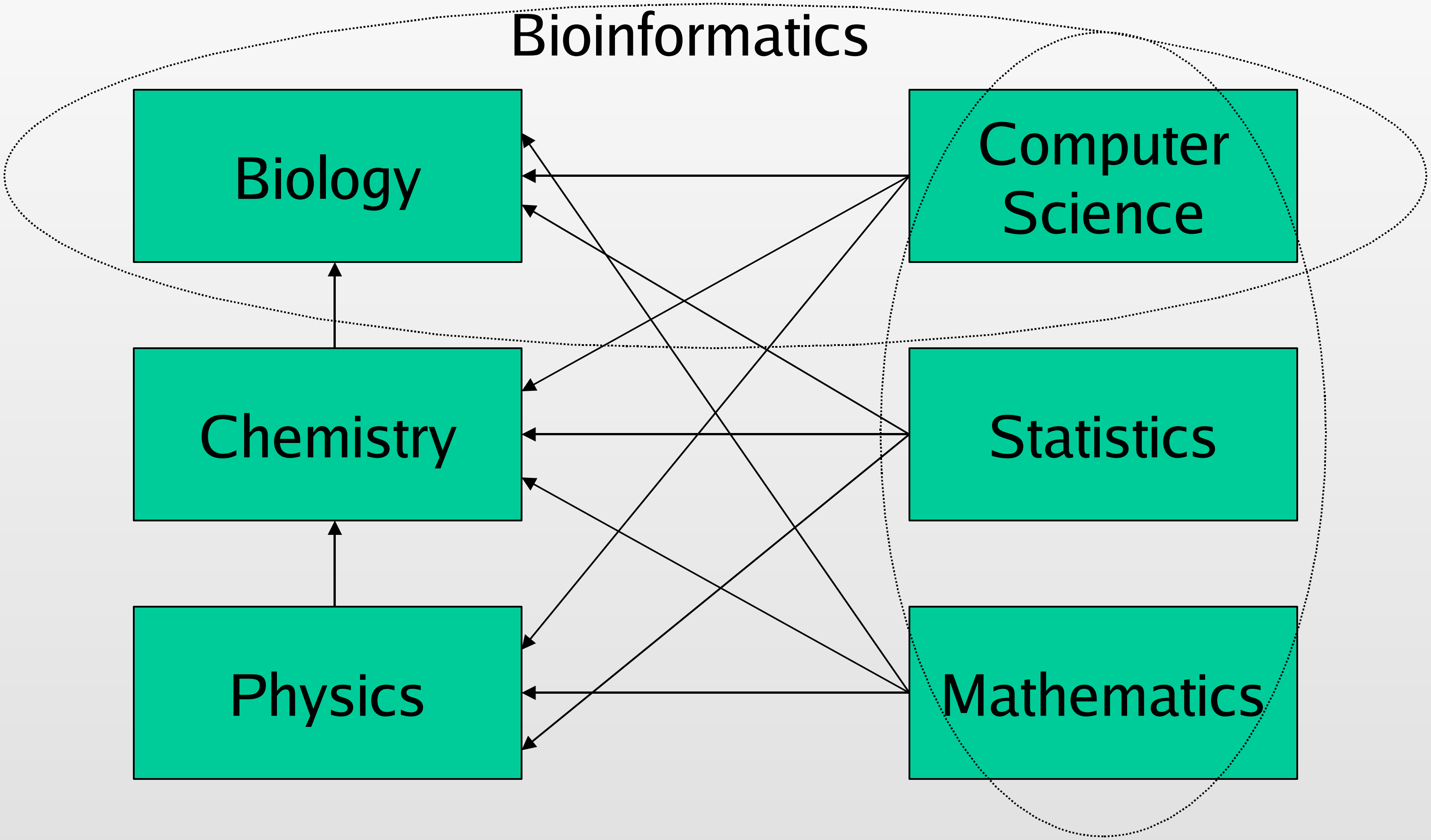
João Meidanis

Based on talk given at SPIRE 2003
Manaus, Brazil

Summary

- Introduction
- Current Challenges in Bioinformatics
  - As seen by Bioscientists
    - Future after Human Genome Project
    - Top Ten Challenges
  - As seen by Computer Scientists
    - Broad Challenges
    - Specialized Challenges
- Perspectives

Dependency among knowledge areas



Applications of Comp Sci to Biology

- Traditionally, number crunching applications (models for biological systems)
- More recently, combinatorial applications, related to DNA and protein sequences, maps, genomes, etc.

- Both Computer Science and Biology deal with very complex systems, e.g., software, cells

How to study complex systems

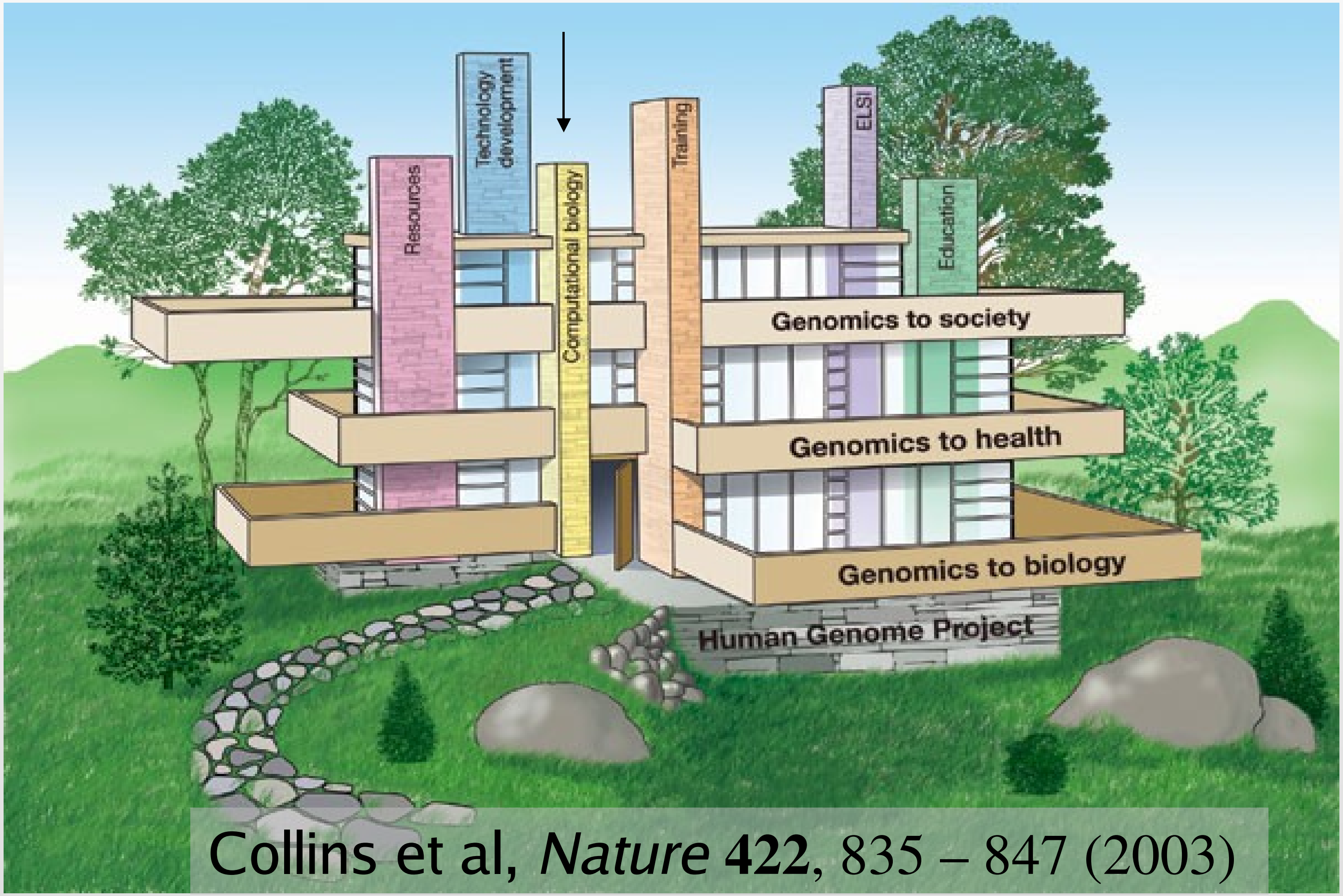
- Study a complex system by taking “projections” or “slices” to focus on one aspect at a time
- Example from CS: **software**: logical view, physical view, development view, etc.
- Example from Biology: **protein**: cellular compartment, biological process, molecular function (as in the Gene Ontology initiative)

Bioinformatics Challenges as seen by Biologists

- Collins et al view of the future of genomics after the Human Genome Project – Computational Biology plays a role
- Top Ten Challenges by Birney, Burge, and Fickett – as we will see, very biologically oriented

The future of Genomics Research

Resources Tech.development



Collins et al, *Nature* **422**, 835 – 847 (2003)

Training ELSI Education

Top Ten Challenges

- Birney (EBI), Burge (MIT), Fickett (GSK), *Genome Technology* 17, Jan 2002

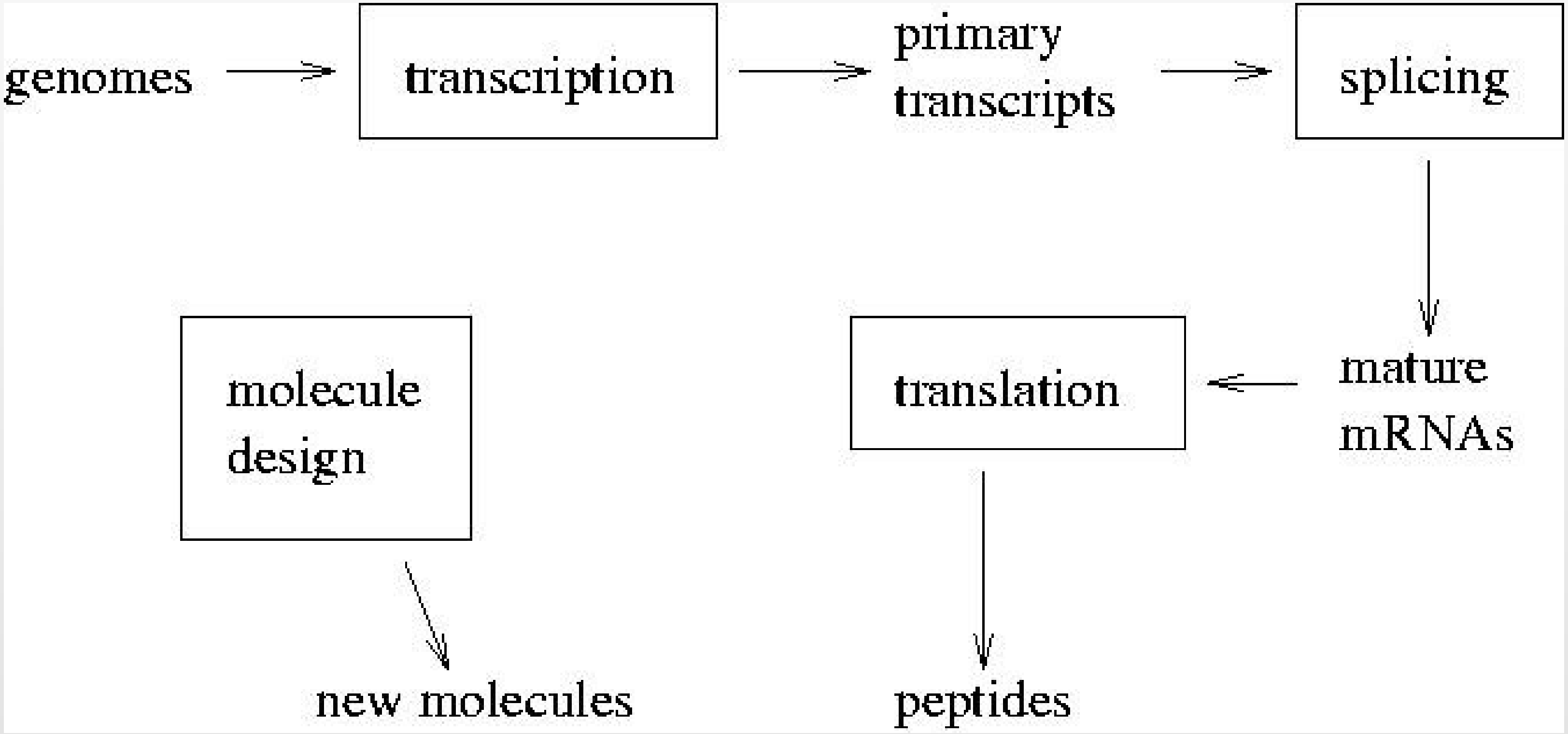
- Predict transcription
- Predict splicing
- Predict signal transduction
- Predict DNA:protein and protein:protein recognition codes
- Predict protein structure

Top Ten Challenges (cont.)

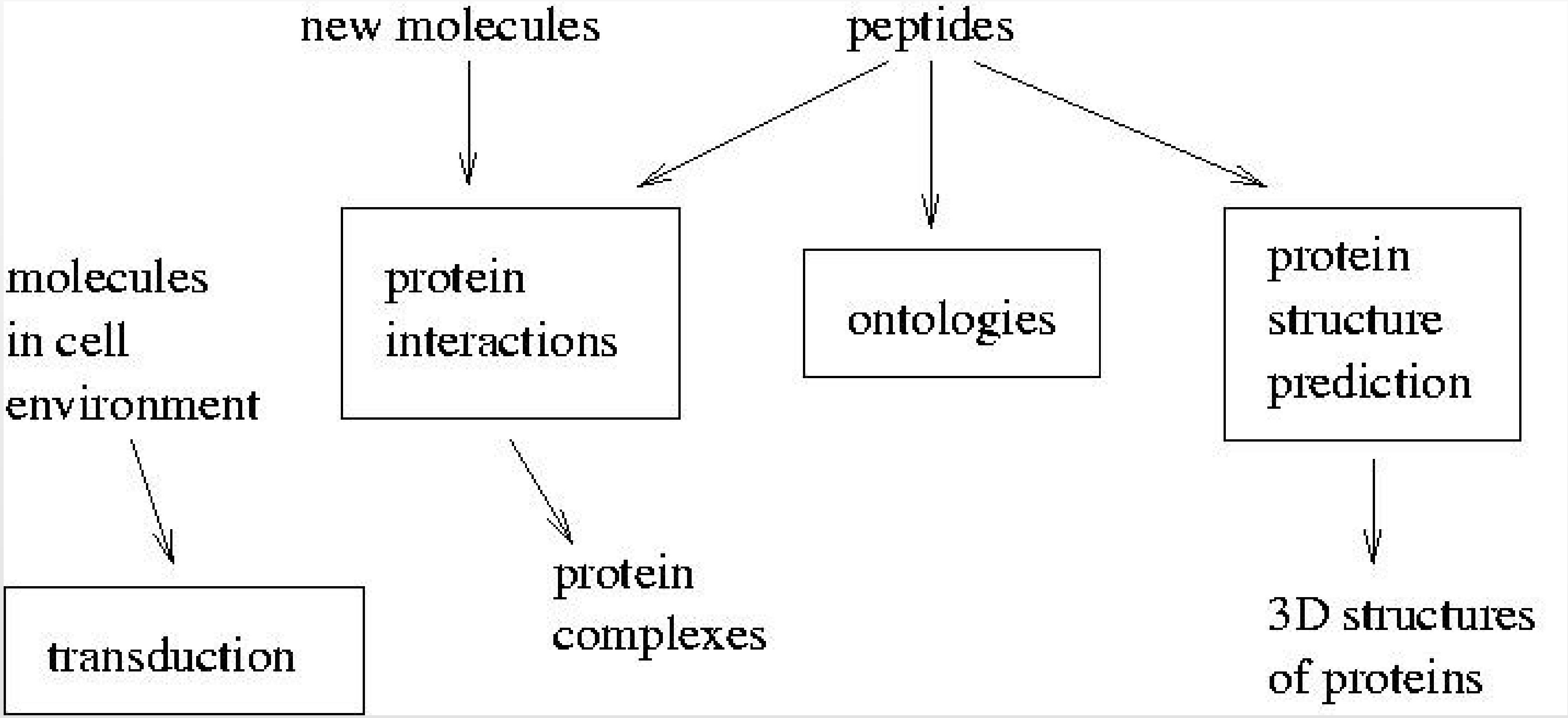
- Birney (EBI), Burge (MIT), Fickett (GSK), *Genome Technology* 17, Jan 2002

- Design small molecule inhibitors of proteins
- Understand protein evolution
- Understand speciation
- Develop effective gene ontologies
- Develop appropriate curricula for bioinformatics education

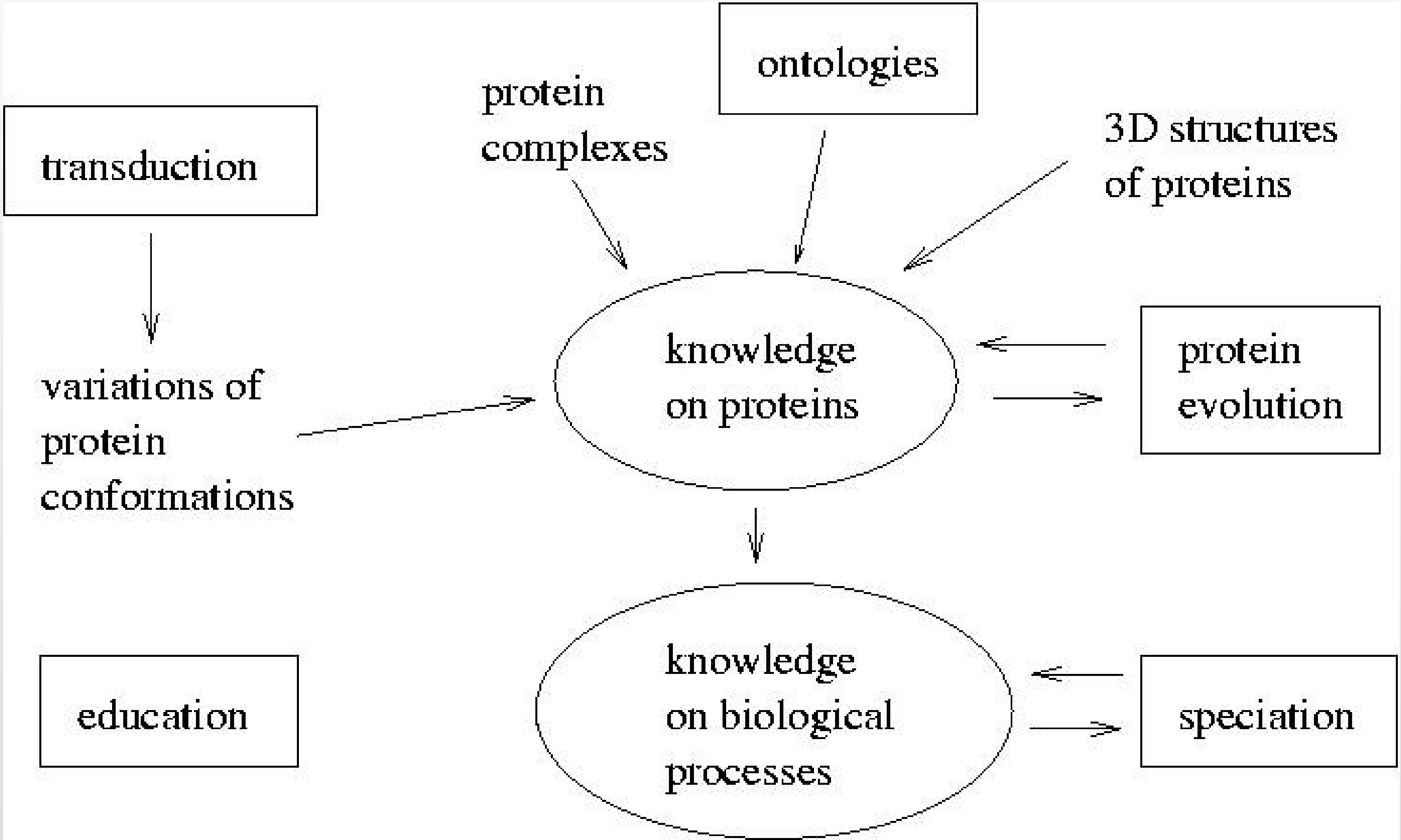
Top Ten, Global View – Part 1



Top Ten, Global View – Part 2


new molecules
peptides
molecules in cell environment
protein interactions
ontologies
protein structure prediction
transduction
protein complexes
3D structures of proteins

Top Ten, Global View – Part 3



transduction

variations of protein conformations

protein complexes

ontologies

3D structures of proteins

knowledge on proteins

protein evolution

education

knowledge on biological processes

speciation

Bioinformatics Challenges as seen by Computer Scientists

- Broad Challenges
  - Information management, paralellism, programability
- Specialized challenges
  - Related to several problems: sequence comparison, fragment assembly and clustering, phylogenetic trees, genome rearrangements and genome comparison, micro-array technology, protein classification

Broad Challenges

- Information management challenge
  - Large sets
  - Semi-structured data
  - Experimental errors
  - Integration of loosely coupled data
- Paralellism challenge
  - Development of expressive control systems for heterogeneous, distributed computing
- Programability
  - Development of higher level languages
  - Programming is still hard and error prone

Limitations of relational databases

- Lack of support for hierarchies
- Changing the schema: all hell breaks loose
- Query language (SQL): it can be challenging and nonintuitive to write an efficient query

Sequence comparison

- Statement of the problem: to find similarities among two or more sequences, usually accompanied by an alignment, highlighting common origin and/or 3D structure
- Many facets of the problem are well understood:
  - Use of dynamic programming
  - Gap-open and gap-extend penalties
  - How to do it using linear space O(m + n)
  - Global, local, semi-global, etc. variants
  - Scoring systems for DNA and protein sequences (e.g., BLOSUM matrices)

Sequence comparison

- But challenges still remain:
  - How to compare very long sequences, e.g., genomes, avoiding the mosaic effect (good regions interspersed with bad regions)
  - One possibility is the use of normalized alignments, where a minimum score per position ratio has to be maintained (Arslan et al, *Bioinformatics* 17:327-337, 2001)
  - How to compare genomic DNA to cDNA sequences
  - Multiple sequence alignment

Fragment assembly

- Statement of the problem: correctly reconstruct a genome (or piece of a genome) from fragments, i.e., contiguous substrings of length ~700

- Facets of the problem that are well understood:
  - Overlap-layout-consensus strategy, its strengths and limitations

Fragment assembly

- Challenges:
  - How to deal with repeats
  - How to use mated pairs and scaffolds
  - Strong dependency on thorough data clean-up
  - Sequencing by hybridization: will it ever be a viable alternative?
  - The Eulerian method: new approach that has not been extensively tested in a production setting (Pevzner et al, *PNAS* 98(17):9748-9753 describes the approach)

EST Clustering

- Statement of the problem: given many samples of mRNAs from the same organism or from closely related organisms, group together in clusters those mRNAs that are related
- Techniques used are similar to those for fragment assembly, but goals are different

EST Clustering

- Challenges:
  - Intended meaning for the cluster: transcript, gene, or gene family
  - How to deal with alternative splicing
  - Strong dependency on thorough data clean-up [Silva and Telles, *Genetics and Molecular Biology* 24(1-4):17-23, 2001 is a good example of thorough clean-up]
  - Recognition of chimeric clones and clusters
  - Separation of paralogs

Physical Mapping

- Statement of the problem: position large, contiguous pieces of a genome in their correct relative location
- Used to be an intermediate step before complete sequencing of a genome
- Now people tend to sequence directly, without mapping first
- Two versions of the problem:
  - Data coming from digestion experiments
  - Data coming from hybridization experiments
- Recent developments
  - PQR trees in almost linear time (Meidanis and Telles, 2003, in preparation)

Phylogenetic trees

- Statement of the problem: construct a tree structure showing the evolution of a group of species from a common ancestor
- Old problem: construction of phylogenetic trees was done using macroscopic characteristics of species before the genomic era
- The area gained momentum with molecular data: differences at the molecular level can be used as characteristics
- It is possible to use distance data originated from sequence comparison as well
- Challenges (just one example):
  - Consensus trees

Genome rearrangements

- Statement of the problem: given two genomes with the same genes, find the minimum number of rearrangement events that lead from one genome to the other
- A crucial observation is that sometimes gene order evolves faster than gene sequence, e.g. in plant mitochondria (Palmer and Herbon, *J. Molecular Evolution* 28:87--97, 1988)
- Possible rearrangement events: reversal, transposition, translocation, fission, fusion, etc.

Genome rearrangements

- The problem was given this precise mathematical formulation recently
- Challenges:
  - To solve the transposition distance problem
  - Combine several events
  - How to deal with gene duplication, gene creation and gene loss (nonconservative comparison)
  - How to compare multiple genomes under rearrangement events

Micro-array Analysis

- Micro-array experiments are one way of measuring the expression pattern of genes, i.e., when and how often a gene is used to produce the corresponding product
- This is not a single bioinformatics problem, but rather requires a collection of problems to be solved in order to design the experiments, gather the results as image files, quantify and normalize the images, and analyze the expression patterns
- It is receiving a tremendous amount of attention

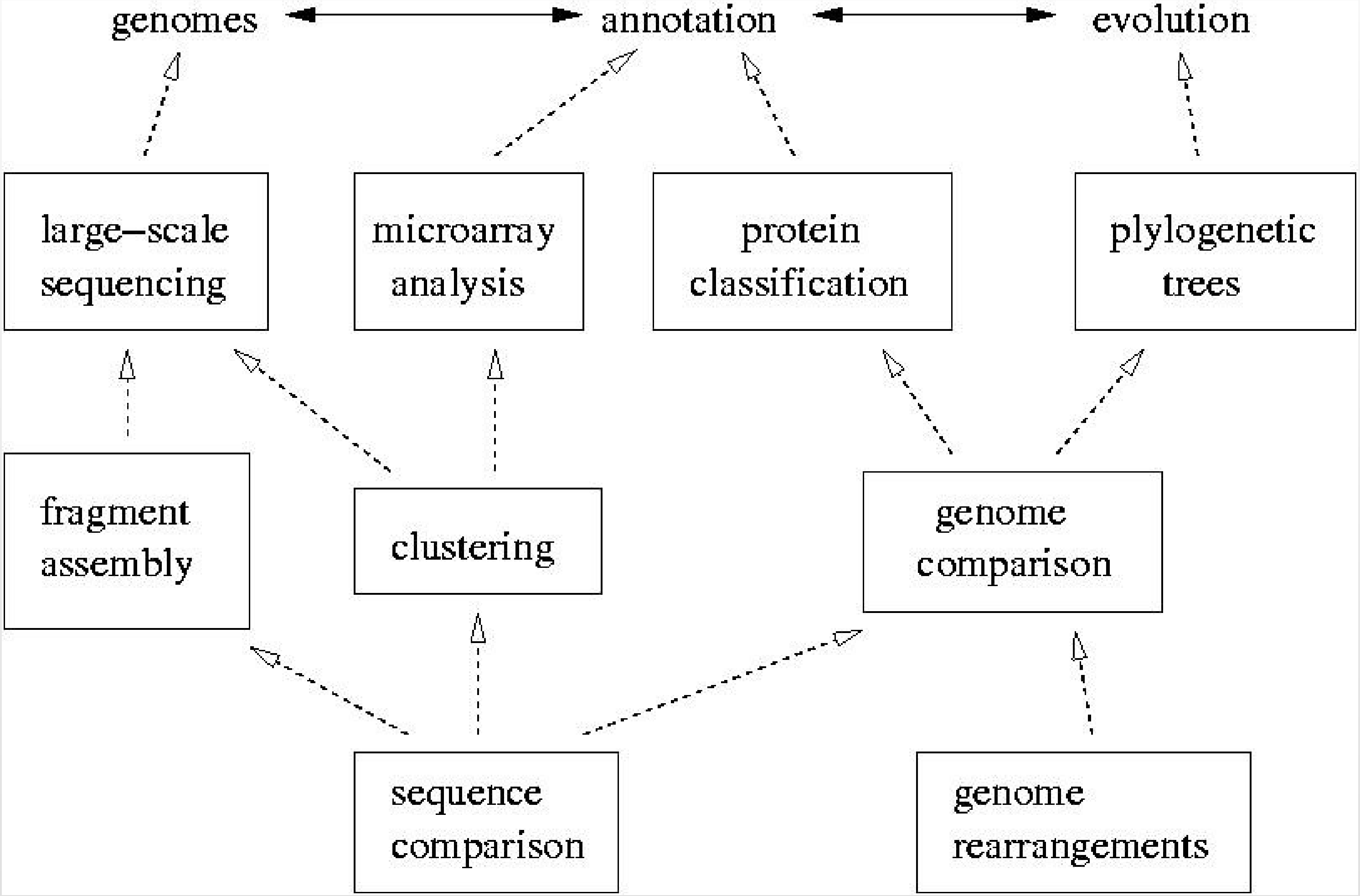
Micro-array Analysis

- Requires strong statistical background
- Challenges:
  - Steps to take to guarantee the reproducibility of results (MIAME - Minimum information about a micro-array experiment - initiative)
  - Clustering algorithms: lots of alternatives, which is the best? (Datta and Datta, *Bioinformatics* 19:459-466, 2003)
  - Data acquisition from images
  - Development of benchmarks (Spellman et al, *Molecular Biology of the Cell* 9:3273-3297, 1998 presented a very influential benchmark set)

Protein Classification

- Statement of the problem: given the sequence of a protein, classify it according to some predefined categorization, usually hierarchical
- The goal is to predict protein function
- There is a huge amount of sequences waiting to be classified
- Challenges
  - Development of automatic classification methods
  - Sequence comparison alone is not sufficient – sequence databases such as GenBank are full of erroneous annotations done by similarity

Specialized Challenges – Global View



Perspectives

- The future of the area will likely include:
  - Formation of larger, interdisciplinary groups
  - Bioscientists and Computer Scientists increasingly understanding both fields
  - Probability and statistics playing an important role
  - Increased quantification
  - Construction of benchmarks