From Sequences to Science - New Computational Perspectives on Old Biological Problems

William R. Pearson

November 17, 2004

From Sequences to Science

- Why Biology isn't like Physics
- Computational Approaches to Biological Problems
  - Sequence Comparison – A success story
  - Gene Prediction – A larger challenge
  - Evolutionary tree reconstruction – Too big, too shallow
- Computational Approaches to Support Discovery
  - False positives are far more expensive than false-negatives
  - Presenting the "shape" of solution space

Why Biology isn't like Physics (diversity)

All science is either physics
or stamp collecting.

–Ernest Rutherford

It is a safe rule to apply that, when a mathematical or philosophical author writes with a misty profundity, he is talking nonsense.

– Alfred North Whitehead, *An Introduction to Mathematics*, 1948

Current sequencing capacity:
- 40 million lanes/year ~ 8 billion bases – Whitehead Inst. (1 bacterial genome/day)

>40 billion bases/year, world-wide
- (1000 bacterial genomes/year; 1 mammalian genome/year)

Foundations of Biology

- Evolution – modern organisms all descended from a common ancestor (Darwin, 1865)
- Survival of the fittest – differences in species result from random mutation and natural selection
- The Central Dogma – genes, the unit of inheritance, are comprised of DNA, which can be replicated, or transcribed into mRNA, which is translated into protein (Avery & McCleod, 1945; Watson & Crick, 1953, et al.)

Insight from Genomes

- 50 – 70% of genes have recognizable homologs in other organisms
- 20 – 30% of genes appear to be “novel”
- fundamental cellular processes – replication, transcription, and metabolic pathways – are largely conserved between very different organisms
- genome size – a major energy consumer in living cells – is not constrained in higher eukaryotes
- many signal transduction pathways are ubiquitous, but different pathways predominate in different organisms

Darwin was right!!!

Common misconceptions

- Natural selection is critical, but not dominant
  - Genome sizes - the C-value paradox
  - The Genetic Code (15 amino acids 50% more efficient)
  - Introns
  - Repeated sequence families
  - Genetic variation, duplications and deletions

Nature operates in the shortest way possible.

– Aristotle 384-322 BCE Physics, Book V

Measures of Size and Complexity

- Number of particles in the universe: ~10^{80}
- Number of possible protein sequences (average length 400 aa): ~20^{400}=10^{520}

The biological world is the result of an evolutionary path, not lowest energy, or optimal cost

From Sequences to Science

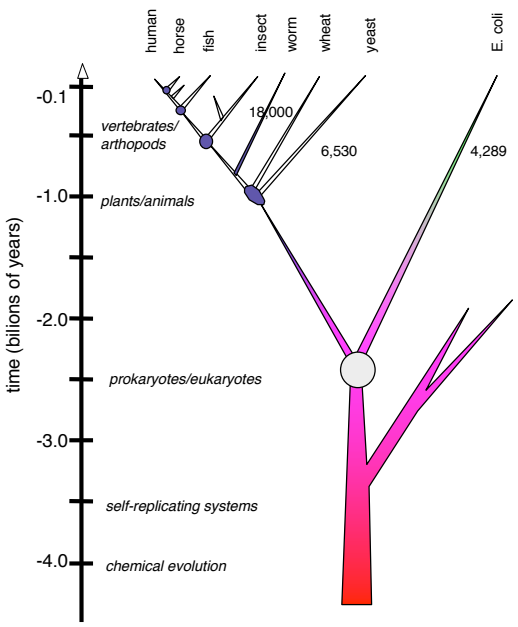
- Why Biology isn't like Physics
- Computational Approaches to Biological Problems
  - Sequence Comparison – A success story
  - Gene Prediction – A larger challenge
  - Evolutionary tree reconstruction – Too big, too shallow
- Computational Approaches to Support Discovery
  - False positives are far more expensive than false-negatives
  - Presenting the "shape" of solution space

Problems in DNA and protein sequence analysis

- Protein analysis:
  - distant relationships
  - sequence to structure
  - sequence/structure to function
  - protein interactions
  - protein identification
- DNA analysis
  - gene finding/splicing
  - promoters/regulatory sites
  - other functional regions
  - DNA and chromatin structure
- Functional/Network Dynamics

Approaches to DNA and protein sequence analysis

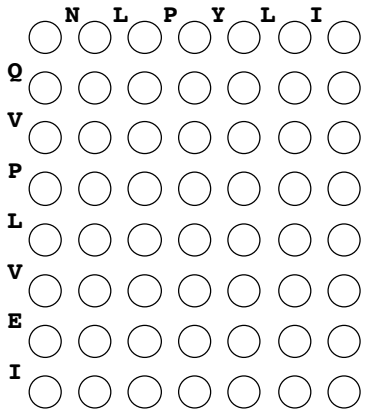
- Evolution based:
  - protein/gene families
  - similarity searching
  - look-back > 2 By
  - accurate statistics
  - common structures
  - common functions?
- Pattern based:
  - promoters
  - functional sites
  - convergence/ not homologs
  - poorly defined/ difficult to identify
  - functional genomics
- Higher order structure?



Human proteins in E. coli

Description	length	E(4400)	%id	E(1700)
IRE1_HUMAN iron-resp. element bind. prot. 1	837			
acnA aconitate hydratase 1	891	1.6e-195	53.4	1.2e-14
PHS_HUMAN glycogen phosphorylase	847			
glgP α-glucan-phosphorylase	815	4.0e-181	49.8	–
MUTA_HUMAN methylmalonyl-coA mutase	750			
sbm chromosome initiation factor	714	1.4e-178	59.3	–
G6PI_HUMAN glucose 6-P isomerase	558			
pgi glucose 6-P isomerase	549	2.2e-164	64.7	2.1e-14
CPSM_HUMAN carbamoyl-P isomerase	1500			
carB carbamoyl-P isomerase	1073	7.2e-162	40.3	2.2e-90
SYV_HUMAN valyl-tRNA synthetase	1263			
valS valyl-tRNA synthetase	951	2.2e-153	40.1	9.5e-72
ODO1_HUMAN 2-oxoglutarate DH E1	1002			
sucA 2-oxoglutarate DH E1	933	2.9e-143	39.1	–
GR75_HUMAN mito. stress-70 prot.	679			
dnaK DNA K protein (HSP70)	638	3.9e-138	60.4	–
DHSA_HUMAN succinate DH	664			
sdhA succinate DH	588	1.2e-126	55.2	1.3e-73
ATPB_HUMAN ATP synth. β-chain	529			
atpD ATP synth. F1 β-subunit	460	3.6e-123	71.7	9.5e-23
BGLR_HUMAN β-glucuronidase	651			
uidA β-D-glucuronidase	603	1.4e-118	45.1	–
SYA_HUMAN alanyl-tRNA synthetase	968			
alaS alanyl-tRNA synthetase	876	4.7e-116	39.1	1.6e-38

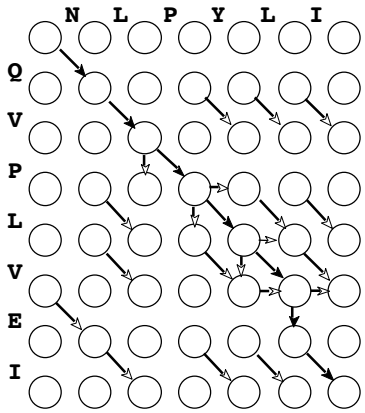
Smith-Waterman



1. score every cell:

$S_{x,y} = max\{$
$S_{x-1,y-1} + match_{xy}$
$S_{x,y-1} - gappen$
$S_{x-1,y} - gappen$
0
$\}$

Smith-Waterman



1. score every cell:

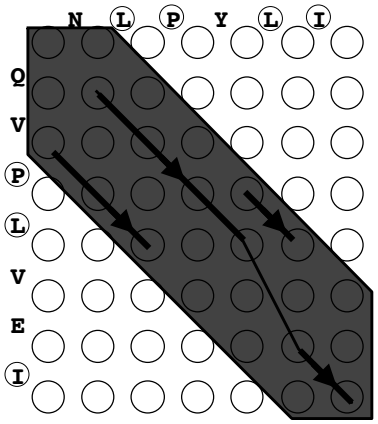
$S_{x,y} = max\{$
$S_{x-1,y-1} + match_{xy}$
$S_{x,y-1} - gappen$
$S_{x-1,y} - gappen$
0
$\}$

2. follow "traceback"

```
NLPYL-I
..: . :
QVPLVEI
```

Outcome: one continuous,optimal gapped alignment

FASTA



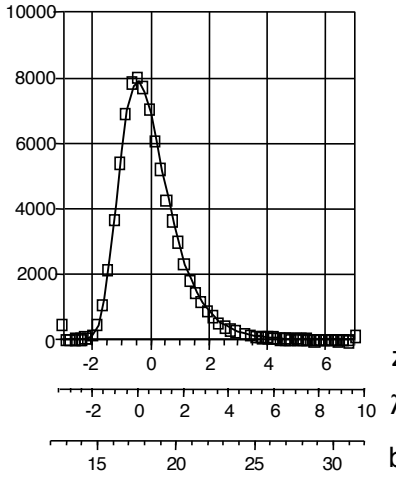
1. Identify identical matches
   (length = *ktup*)
2. Extend along diagonal
   (local maximum)
3. Join diagonal segments (DP)
   (maintain linearity)
   (optimal sum score)

4. Banded Smith-Waterman

```
NLPYL-I
..: . :
QVPLVEI
```

Outcome: one continuous, near-optimal gapped alignment

Extreme value distribution



$S' = \lambda S_{raw} - \ln K\, m\, n$

$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$

$P(S'>x) = 1 - \exp(-e^{-x})$

$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$

$E(S'>x \mid D) = P\, D$

$P(B \text{ bits}) = m\, n\, 2^{-B}$

$P(40 \text{ bits}) = 1.5 \times 10^{-7}$

$E(40 \mid D=4000) = 6 \times 10^{-4}$

$E(40 \mid D=2E6) = 0.3$

Inferring Homology from Statistical Significance

- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM,* then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

Challenging Computations – Gene Finding

Table 7. Sensitivity and specificity of Otto and Genscan.
Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)	0.604	0.884
Genscan	0.501	0.633

* Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. Refers to those annotations produced by supplying all available evidence to Genscan.

Science (2001) 29:1304-51

In fly, *ab initio* methods can correctly predict around 90% of individual exons and can correctly predict all coding exons of a gene in about 40% of cases. For human, the comparable figures are only about 70% and 20%, respectively. These estimates may be optimistic, owing to the design of the tests used.

Nature (2001) 409:860-921

Improving Gene Finding – Evolution

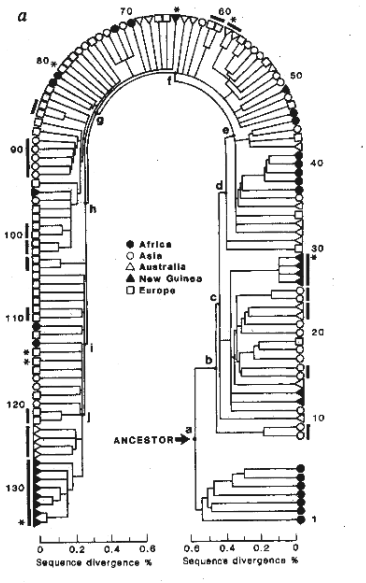
Korf I, Flicek P, Duan D, Brent MR (2001) Bioinformatics 17 Suppl 1:S140-148
Integrating genomic homology into gene structure prediction.

TWINSCAN is a new gene-structure prediction system that directly extends the probability model of GENSCAN, allowing it to exploit homology between two related genomes. Separate probability models are used for conservation in exons, introns, splice sites, and UTRs, reflecting the differences among their patterns of evolutionary conservation. TWINSCAN is specifically designed for the analysis of high-throughput genomic sequences containing an unknown number of genes. In experiments on high-throughput mouse sequences, using homologous sequences from the human genome, TWINSCAN shows notable improvement over GENSCAN in exon sensitivity and specificity and dramatic improvement in exact gene sensitivity and specificity. This improvement can be attributed entirely to modeling the patterns of evolutionary conservation in genomic sequence.

Mitochrondrial “Eve”

Cann RL, Stoneking M, Wilson AC. (1987) “Mitochondrial DNA and human evolution.” Nature. 325:31-36.

Mitochondrial DNAs from 147 people, drawn from five geographic populations have been analysed by restriction mapping. All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa. All the populations examined except the African population have multiple origins, implying that each area was colonised repeatedly.



Pearson WR, Robins G, Zhang T. (1999) Mol Biol Evol. 16:806-816.
Generalized neighbor-joining: more reliable phylogenetic tree reconstruction

We have developed a phylogenetic tree reconstruction method that detects and reports multiple topologically distant low-cost solutions. Our method is a generalization of the neighbor-joining method of Saitou and Nei and affords a more thorough sampling of the solution space by keeping track of multiple partial solutions during its execution.

From Sequences to Science

- Why Biology isn't like Physics
- Computational Approaches to Biological Problems
  - Sequence Comparison – A success story
  - Gene Prediction – A larger challenge
  - Evolutionary tree reconstruction – Too big, too shallow
- Computational Approaches to Support Discovery
  - False positives are far more expensive than false-negatives
  - Presenting the "shape" of solution space

From Sequences to Science

Computational Approaches to Support Discovery

- False positives are far more expensive than false-negatives
  - Today, biological discoveries remain "individual" – once a gene has been identified using statistical or genomics approaches, it is "confirmed" by identifying the specific mutations associated with a disease AND proposing a mechanism
  - Conformation is very highly accurate, and cost-effective for dozens, but not hundreds, of false-positives
  - There are 20,000 (?) genes, and $10^{?}$ interactions, so 10^{-3} false positive rates are high.
  - Biologists value their creativity, and some enjoy seeing computer predictions that are nonsensical.
- Presenting the "shape" of solution space
  - Biology is a search for alternate explanations with the expectation that only a few are correct; most are misleading
  - Explanations that are very "different" yet explain similar phenomena are of great interest.

There are more things in heaven and earth,
Horatio, Than are dreamt of in your philosophy.
– William Shakespeare, Hamlet, 1601, (LA Story)