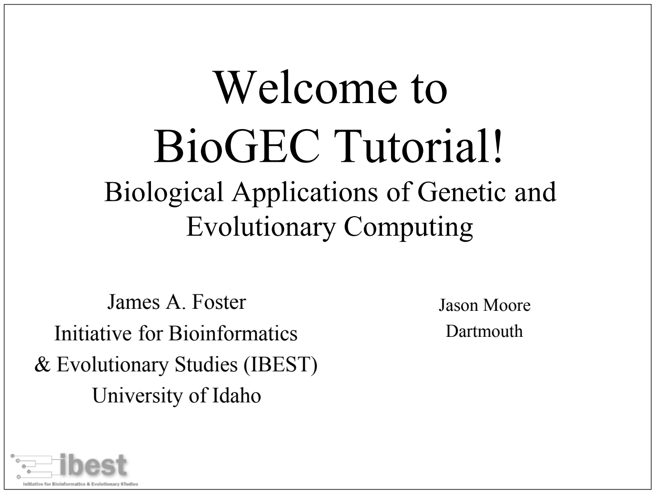
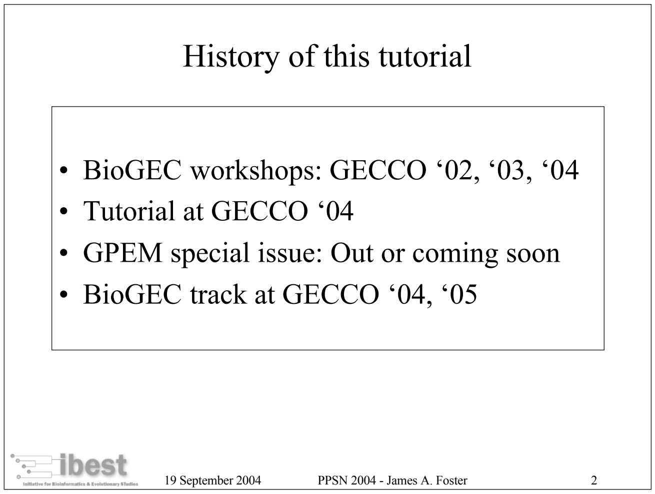


Welcome to BioGEC Tutorial!

Biological Applications of Genetic and Evolutionary Computing

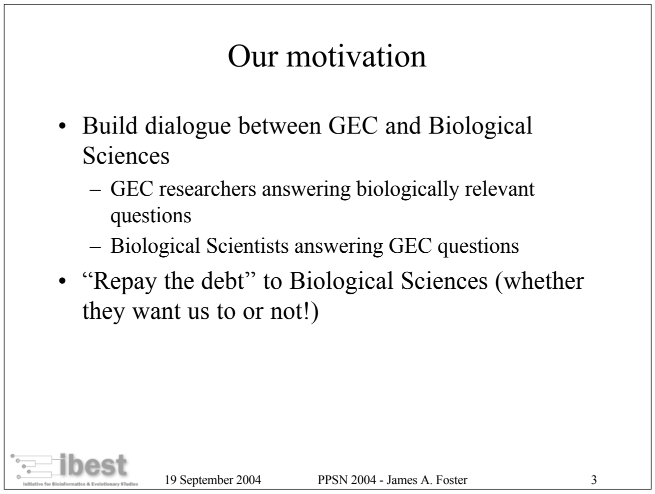
James A. Foster
Initiative for Bioinformatics
& Evolutionary Studies (IBEST)
University of Idaho

Jason Moore
Dartmouth



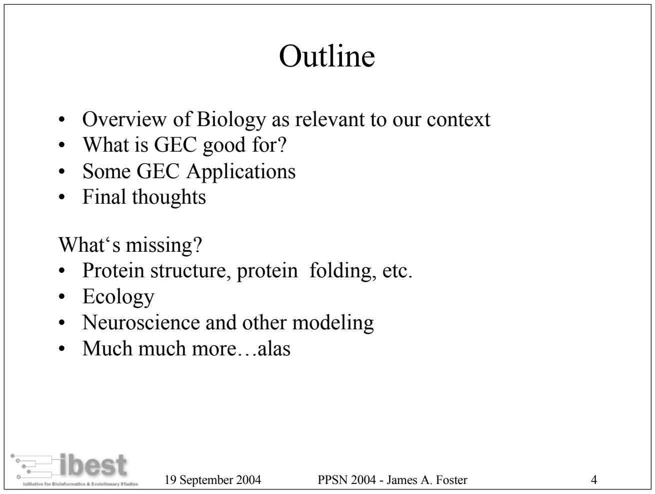
History of this tutorial

- BioGEC workshops: GECCO ‘02, ‘03, ‘04
- Tutorial at GECCO ‘04
- GPEM special issue: Out or coming soon
- BioGEC track at GECCO ‘04, ‘05



Our motivation

- Build dialogue between GEC and Biological Sciences
  - GEC researchers answering biologically relevant questions
  - Biological Scientists answering GEC questions
- “Repay the debt” to Biological Sciences (whether they want us to or not!)



Outline

- Overview of Biology as relevant to our context
- What is GEC good for?
- Some GEC Applications
- Final thoughts

What‘s missing?

- Protein structure, protein folding, etc.
- Ecology
- Neuroscience and other modeling
- Much much more…alas

PART I

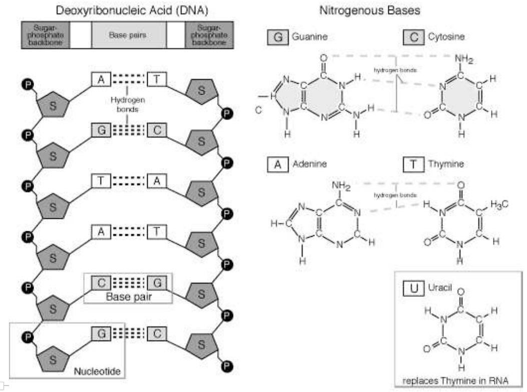
(Really) Basic Biology

Overview

***Biology*: How living things do what they do, and how they came to do it.**

Sequences	Molecules: DNA, RNA, proteins
Structure	Organized into 3D complexes
Function	Interacting with each other
Ecology	Other organisms & environments
Evolution	Changing over time

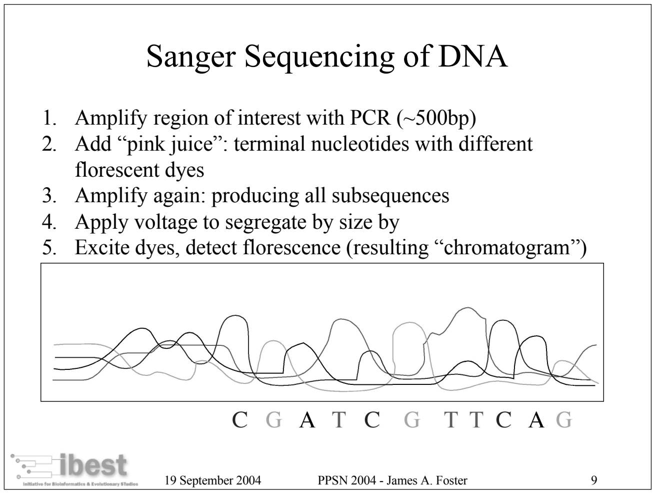
Nucleic Acids: binding with a twist



Copyright 1999 Access Excellence @ the National Health Museum.

Types of DNA

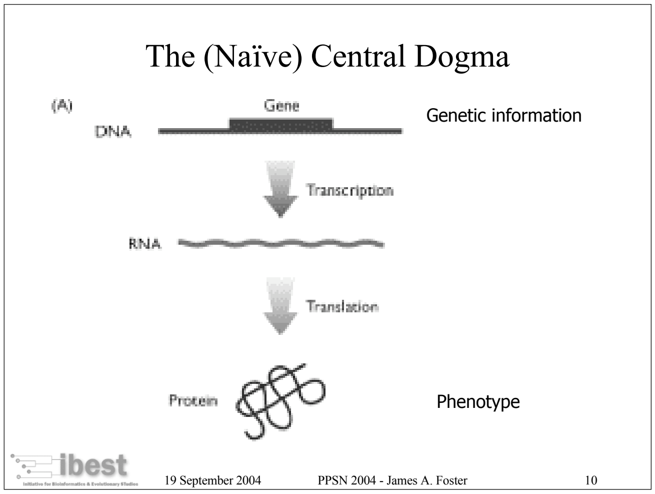
- Genomic
  - Genic: codes for proteins for the host (30K in humans, 100K proteins)
  - Non-Genic (95% of humans)
    - RNA coding
    - Regulatory
    - Mobile elements (over 40% of humans)
    - Endogenous retroviruses
- Non-genomic
  - Organelles (mitochondria, chloroplasts)
  - Plasmids



Sanger Sequencing of DNA

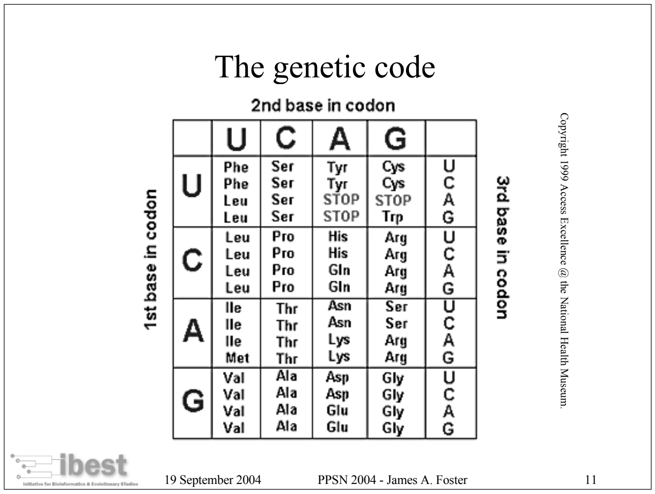
1. Amplify region of interest with PCR (~500bp)
2. Add “pink juice”: terminal nucleotides with different florescent dyes
3. Amplify again: producing all subsequences
4. Apply voltage to segregate by size by
5. Excite dyes, detect florescence (resulting “chromatogram”)

C G A T C G T T C A G



The (Naïve) Central Dogma

(A)
Gene
DNA
Genetic information
Transcription
RNA
Translation
Protein
Phenotype

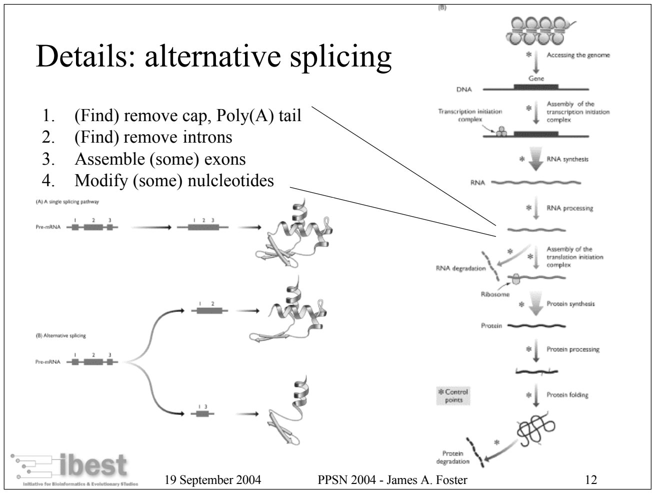


The genetic code

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

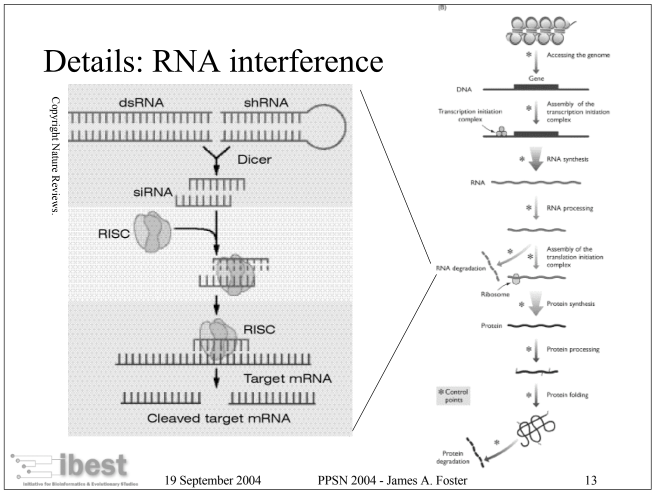
Copyright 1999 Access Excellence @ the National Health Museum.



Details: alternative splicing

1. (Find) remove cap, Poly(A) tail
2. (Find) remove introns
3. Assemble (some) exons
4. Modify (some) nulcleotides

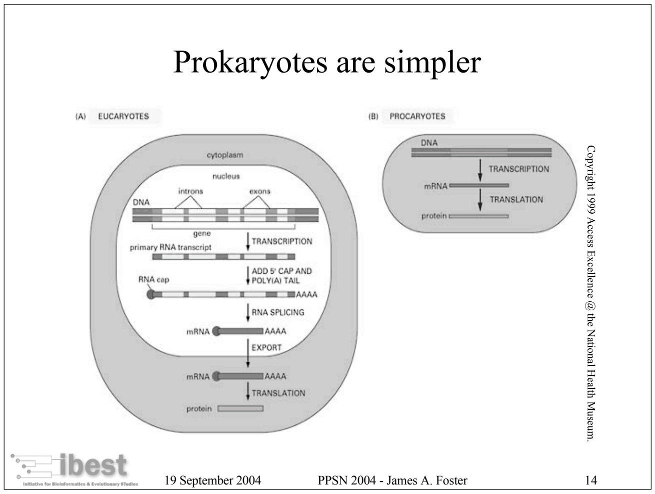
(A) A single splicing pathway
Pre-mRNA
1 2 3
1 2 3
(B) Alternative splicing
Pre-mRNA
1 2 3
1 2
1 3
(B)
Accessing the genome
Gene
DNA
Transcription initiation complex
Assembly of the transcription initiation complex
RNA synthesis
RNA
RNA processing
Assembly of the translation initiation complex
RNA degradation
Ribosome
Protein synthesis
Protein
Protein processing
✻ Control points
Protein folding
Protein degradation



Details: RNA interference

Copyright Nature Reviews.

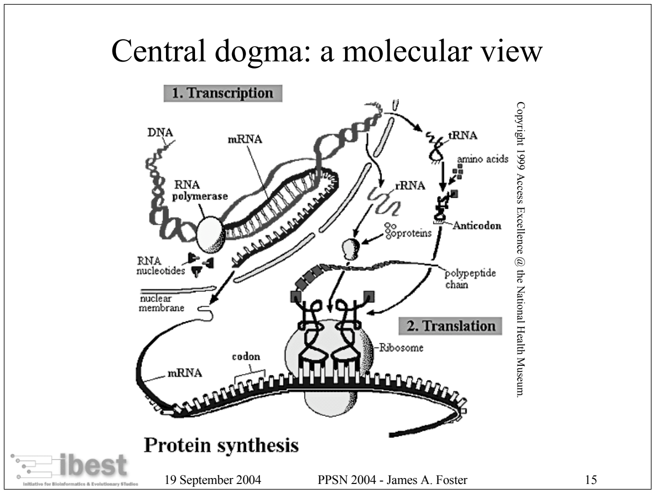
19 September 2004 PPSN 2004 - James A. Foster 13



Prokaryotes are simpler

Copyright 1999 Access Excellence @ the National Health Museum.

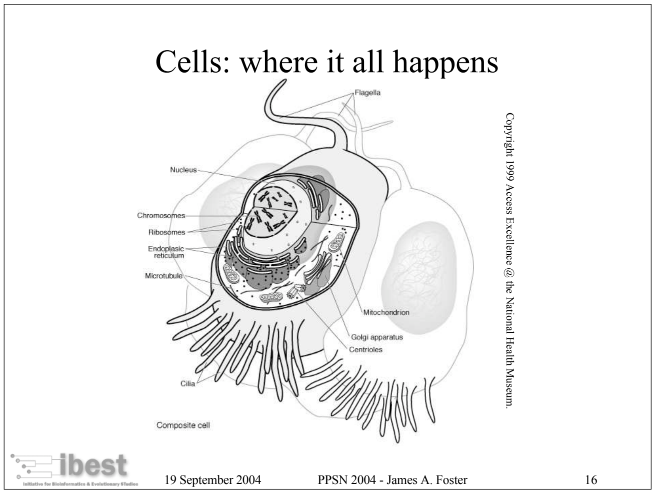
19 September 2004 PPSN 2004 - James A. Foster 14



Central dogma: a molecular view

Copyright 1999 Access Excellence @ the National Health Museum.

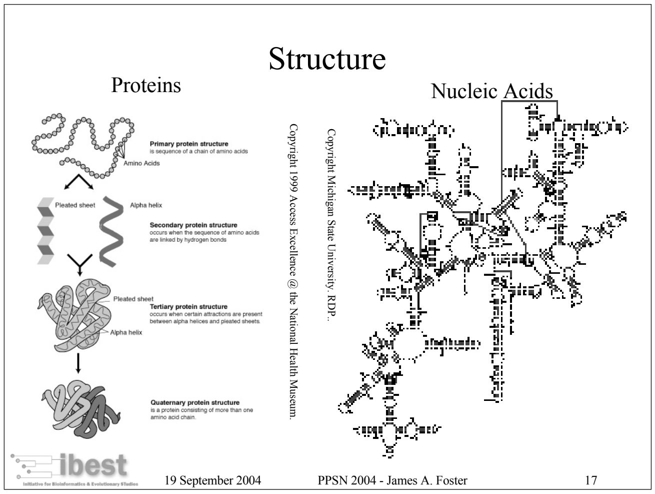
19 September 2004 PPSN 2004 - James A. Foster 15



Cells: where it all happens

Copyright 1999 Access Excellence @ the National Health Museum.

19 September 2004 PPSN 2004 - James A. Foster 16



Structure

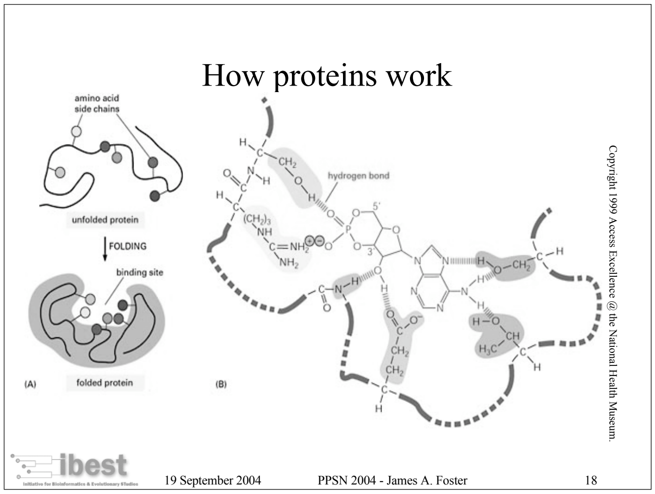
Proteins

Nucleic Acids

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

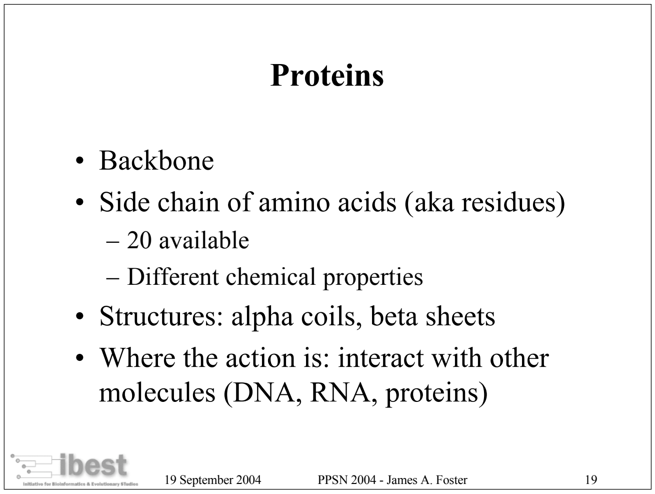
Copyright 1999 Access Excellence @ the National Health Museum.

Copyright Michigan State University. RDP..



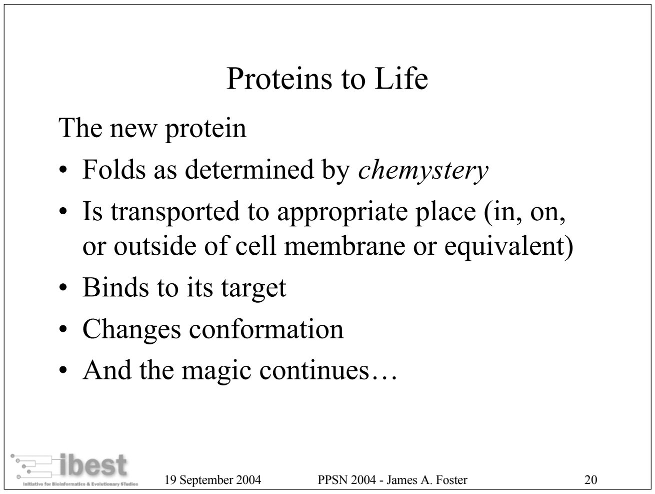
How proteins work

Copyright 1999 Access Excellence @ the National Health Museum.



Proteins

- Backbone
- Side chain of amino acids (aka residues)
  - 20 available
  - Different chemical properties
- Structures: alpha coils, beta sheets
- Where the action is: interact with other molecules (DNA, RNA, proteins)



Proteins to Life

The new protein

- Folds as determined by *chemystery*
- Is transported to appropriate place (in, on, or outside of cell membrane or equivalent)
- Binds to its target
- Changes conformation
- And the magic continues…

Ecology

Studies the distribution, abundance, interactions of organisms & environment

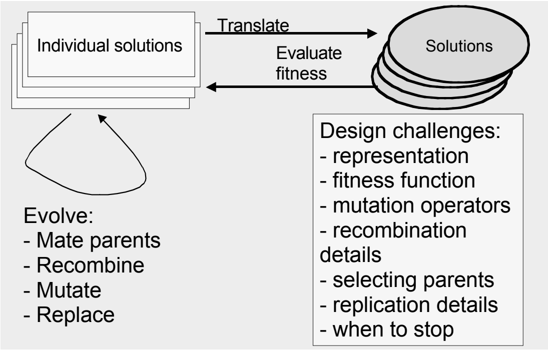
- Why are organisms distributed the way they are (modeling)?
- How does speciation happen (evolution)?
- Define & quantify inter-species interactions (systems science/networks)?
- What is role of randomness in biological systems (modeling)?

Evolution: accumulating changes

- Errors happen:
  - While transcribing: misreads, slipping, breaks, exon duplication, gene duplication
  - While sitting in the cell: rearrangements, chromosomal duplication
  - From the outside: viruses, mobile elements
- Isolated populations get different errors
- We observe only those differences that persist
  - Because they actually help (selection)
  - Or just because (neutral evolution)

Part II
Why use evolutionary algorithms to study biology?

What is Evolutionary Computing?



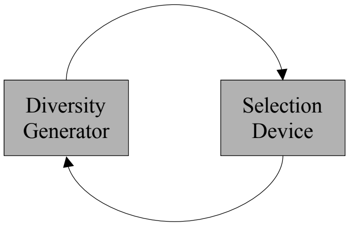
Evolve:
- Mate parents
- Recombine
- Mutate
- Replace

Design challenges:
- representation
- fitness function
- mutation operators
- recombination details
- selecting parents
- replication details
- when to stop

What is Genetic and Evolutionary Computing?

Variation – selection loop

- Iterative discovery & exploitation of objects built from parts correlated with algorithmic objectives



What is GEC good for?

Problems that:

- Are parameter rich
- Are ill-posed: “A problem is well-posed when a solution *exists*, is *unique*, and *depends continuously on the initial data*. It is *ill-posed* when if fails to satisfy *at least one* of these criteria. (Hadamard)“
  - You know what you want, but not how to get it
  - You can compare solutions, but not characterize “the best one”
  - Modeling is inappropriate or premature
- Are noisy
- Have vast, rugged search spaces
- Have non-linear features
- Note: “NP complete” is not enough of a reason

Other advantages of GEC

- GEC can surprise you
  - Lacks strong bias of model-based algorithms
  - Building block sampling is parameter selection
  - Generates hypotheses
  - Less risk of building in biases
- Possibilities for post-run analyses
  - Populations provide samples that can be used for statistical analysis
  - Solutions may show features that can be modeled and used for better algorithms later

Summary and Outlook

Problems in Biology often are often:

- High-dimensional
- Ill-posed
- Based on noisy data
- Non-linear
- Too complicated or poorly understood for modeling

Powerful techniques such as GEC will make it easier to move from descriptive to explicative and predictive models of biological systems

Part III
Some Applications

Case Study
Multiple sequence alignment

The Problem

How are these sequences related???

1	AAGTTTTCCTGGTTCAGTATCCCTAGACC
2	AAGTTTTCGTGGATCACTATCCCTAGAC
3	AAGTTTTCGTGAGTCGATATCCCTAGACT
4	AGTTTTCGTCGGTCGATTATCCCTAGAC
5	TTTCCTGGCTCAGTCCTAATCCCTAGA
6	AAGTTTCCTGGATCAGAATCCCTAGACC
7	AAGTTTCCAGGCTCAGTATCCCTAGACC
8	AAGTTTCCAGGCTAGTATCCCTAGACC

Account for: insertions, deletions, duplications, rearrangements, changes -- conserved through evolution

Goal: Determine related characters

	Alignment (color key: ok, maybe, stretch, beats me)
	AAGTTTTCC-TGGNGTCCA-GTAATCCCTAGACC
1	AAGTTTTCC-TGGT-T-CA-GT-ATCCCTAGACC
2	AAGTTTT-CGT-GGAT-CA-CT-ATCCCTAGA-C
3	AAGTTTT-CGT-GAGTCGA--T-ATCCCTAGACT
4	-AGTTTT-CGTCG-GTCGAT-T-ATCCCTAGA-C
5	----TTTCC-TGGCTCAGTCCTAATCCCTAGA--
6	AAG-TTTCC-TGG-AT-CA--GAATCCCTAGACC
7	AAG-TTTCC-AGGC-T-CA-GT-ATCCCTAGACC
8	AAG-TTT-C-CAG-G-CTA-GT-ATCCCTAGACC

Why should I care about MSA?

- First step to
  - Phylogenetic analysis
  - Database lookup (Blast and PsiBlast)
  - Discovery of important binding sites, motifs, etc.
  - Classification of molecules
  - Etc. (a big “etc.”)

Possible algorithms

- Try all possible alignments
- Iterative: put in gaps, move them around
- Stochastic: Look for patterns, start with them
- Heuristic: do something that “makes sense”

Bottom line: no algorithm is likely that is both efficient and accurate--MSA is “NP hard”

GEC approaches

- Placement of gaps
  - Directly (Zhang & Wong; Shyu et al.)
  - With heuristic rules (Notredame (SAGA))
  - To improve alignments gathered in other ways (Chellapilla & Fogel)
- Evolving guide trees (Sheneman et al. (Evalyn))

SAGA: Seq. Alignment with GAs

Notredame & Higgins ‘96

- Fitness:
  - *Sum of Pairs scoring*: maximize sum of scores for each pair of characters in each column, using fixed scoring matrix, affine gap model
- 22 Operators, dynamically scheduled
  - Mutations (20), applied with evolving rates
    - 16 ways to shift gaps left and right
    - Gap insertion into estimated homologous clades, hillclimbing
    - Block searching
    - Local rearrangement
  - Crossovers (2)
    - One point with gaps as needed
    - Uniform between conserved columns

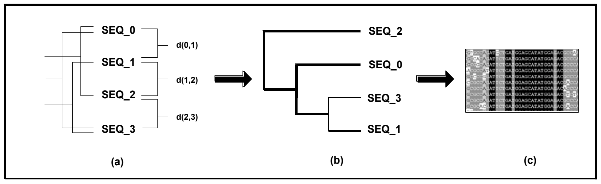
Testing SAGA

- Compared to Clustal W
  - 9 “small” test sets: 4-8 sequences, 60-280 characters
  - 3 “larger” sets: 9, 12, 15, 32 sequences
  - (later) compared to BAliBase test suite
- Higher scoring alignments, but very slow
- Noted that:
  - Sum of Pairs may not be ideal
  - conserved columns measures of “consistency”

Progressive MSA (e.g. Clustal)

Build a *guide tree* (b) (neighbor joining or UPGMA) from pairwise distances (a)

Align pairs of sequences from bottom up, using dynamic programming (c)

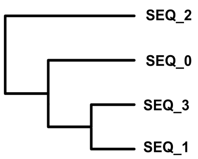


EVALYN: EC for progressive MSA

Idea: discover better guide trees by combining “good” features through evolution

Result: superior alignments (measured by sum of pairs scoring), faster algorithm for large numbers of taxa

Evolve this!



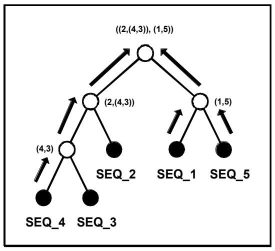
Initialization & Representation

- *Initial population*: randomly generated rooted, bifurcating trees, with each taxon labeling exactly one leaf
- *Representation of individuals*: bifurcating trees with each taxon uniquely present in leaves

Fitness

Build alignment progressively using individual guide tree

- *Sum of pairs*: for each column, add score for aligning each pair of characters; add column scores



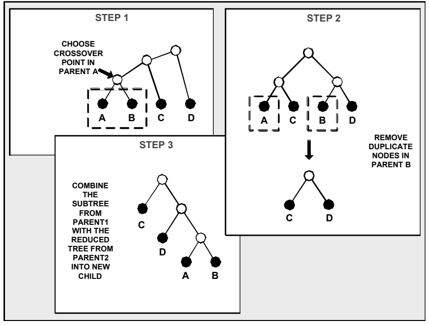
Progressively Aligning Sequences Using a Guide Tree

Selection & Replacement

- Each individual (genotype) produces an alignment (phenotype) with a score (fitness)
- Individuals selected for reproduction with recombination & mutation proportionally to fitness
- Children replace individuals (guide trees) with lower fitness

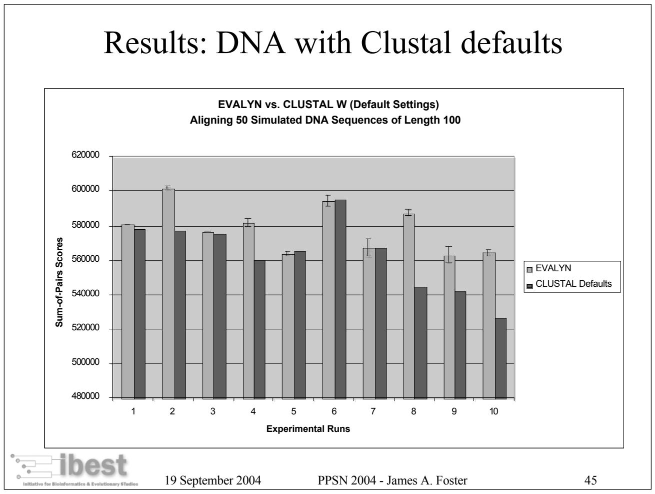
Recombination

- Branch swapping *between* two trees

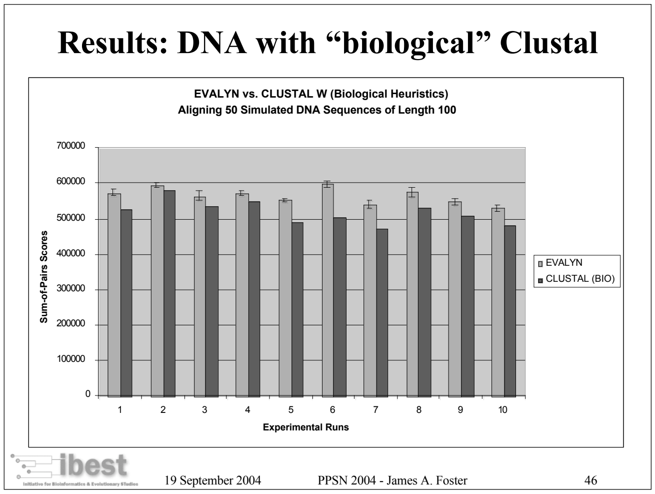

STEP 1
CHOOSE CROSSOVER POINT IN PARENT A
A
B
C
D
STEP 2
A
C
B
D
REMOVE DUPLICATE NODES IN PARENT B
C
D
STEP 3
COMBINE THE SUBTREE FROM PARENT1 WITH THE REDUCED TREE FROM PARENT2 INTO NEW CHILD
C
D
A
B

Testing Evalyn: Setup 1

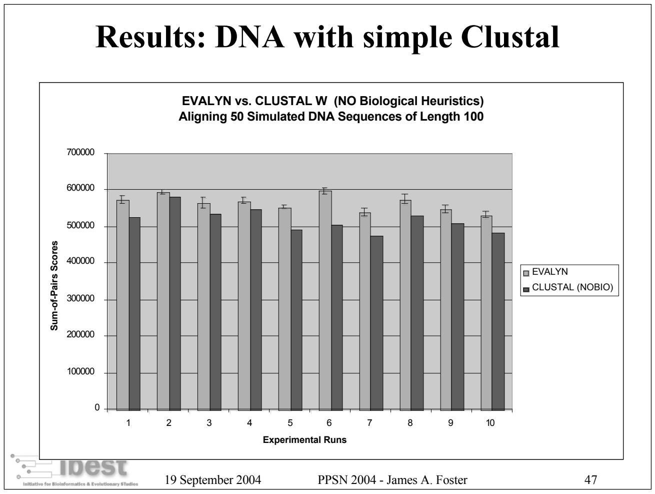
- Simulate 50 variations of a fixed DNA sequence of 100 bp under Jukes-Cantor model (note: star toplogy)
- Align with EVALYN and CLUSTAL (default/adaptive, biological, basic parameters)
- Use CLUSTAL to score alignments
- EVALYN Parameters:
  - Population Size = 500
  - Iterations = 25,000
  - Mutation Rate = 0.01
  - Match = 2.0, Mismatch = 0.0, Gap Open = -10, Gap Extend = -.1



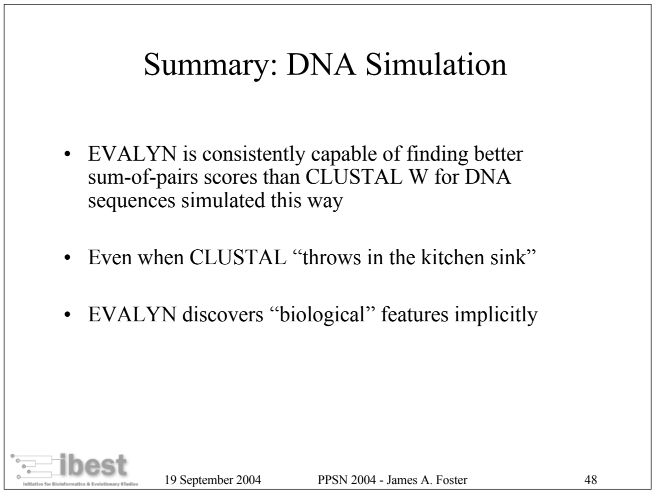
Results: DNA with Clustal defaults



Results: DNA with “biological” Clustal

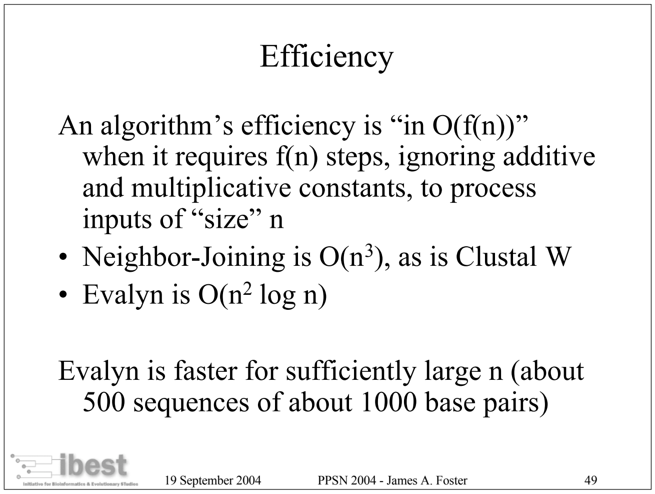


Results: DNA with simple Clustal



Summary: DNA Simulation

- EVALYN is consistently capable of finding better sum-of-pairs scores than CLUSTAL W for DNA sequences simulated this way
- Even when CLUSTAL “throws in the kitchen sink”
- EVALYN discovers “biological” features implicitly

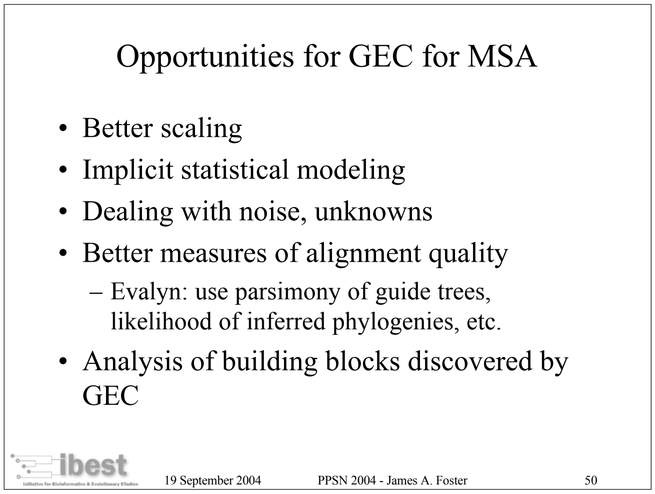


Efficiency

An algorithm's efficiency is "in O(f(n))" when it requires f(n) steps, ignoring additive and multiplicative constants, to process inputs of "size" n

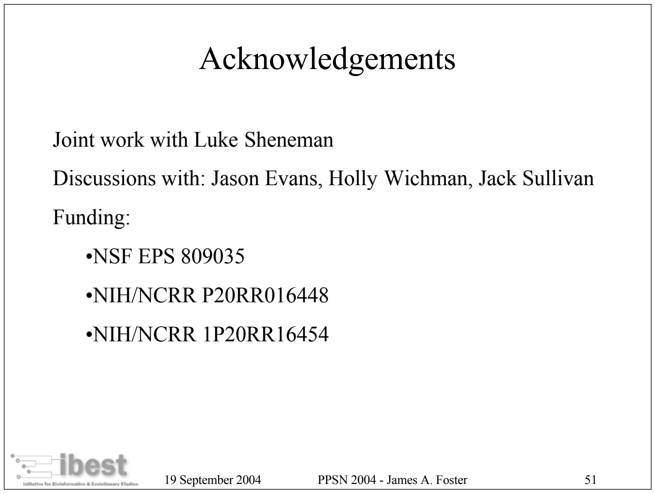
- Neighbor-Joining is $O(n^3)$, as is Clustal W
- Evalyn is $O(n^2 \log n)$

Evalyn is faster for sufficiently large n (about 500 sequences of about 1000 base pairs)



Opportunities for GEC for MSA

- Better scaling
- Implicit statistical modeling
- Dealing with noise, unknowns
- Better measures of alignment quality
  - Evalyn: use parsimony of guide trees, likelihood of inferred phylogenies, etc.
- Analysis of building blocks discovered by GEC



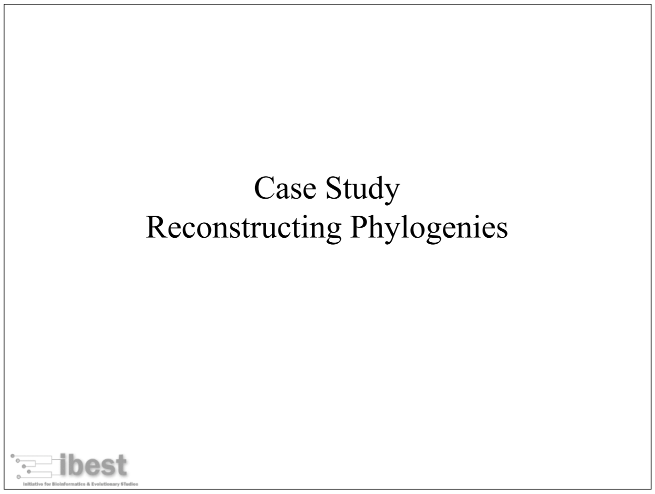
Acknowledgements

Joint work with Luke Sheneman

Discussions with: Jason Evans, Holly Wichman, Jack Sullivan

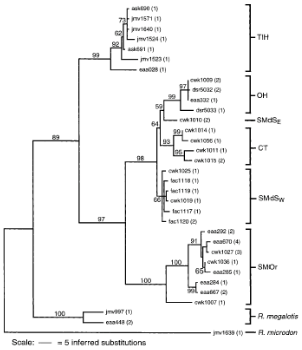
Funding:

- •NSF EPS 809035
- •NIH/NCRR P20RR016448
- •NIH/NCRR 1P20RR16454



Case Study
Reconstructing Phylogenies

The problem



The Problem

- Given: characters related by evolution (an MSA)
- Find: correct (or plausible) evolutionary history that produced them

- Applications:
  - Finding related genes or gene products
  - Resolving taxonomies

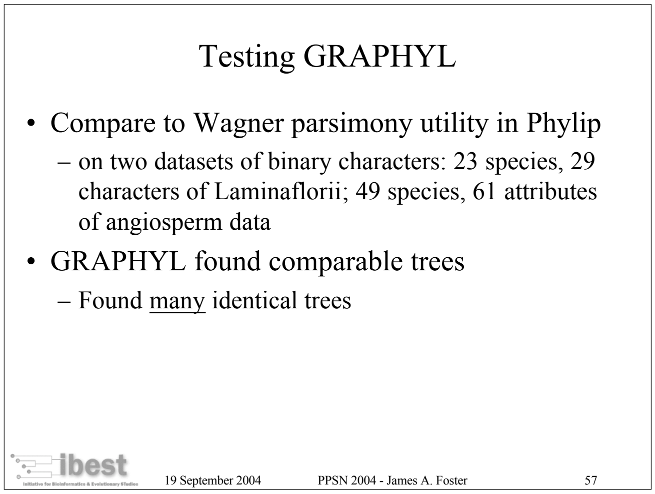
- Challenges
  - Vast search spaces
  - Parameter rich statistical modeling

Current Approaches

- Distance based algorithms
  - Maximum parsimony: minimize changes required to explain data
  - Clustering: group sequences by similarity
- Model based algorithms
  - Maximum likelihood: find tree that maximizes probability of data, given model of evolution

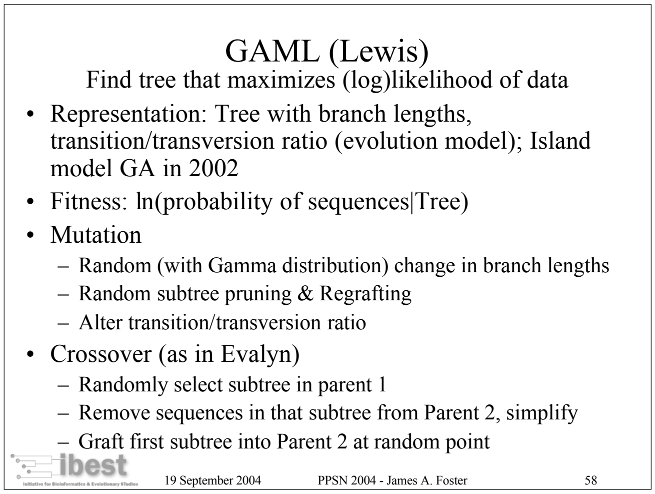
GRAPHYL (Congdon)

- Find tree that minimizes number of changes along branches (maximizes parsimony)
  - Representation: canonical form determined by input dataset
  - Fitness: parsimony of trees
  - Crossover:
    - Select subtree in parent 1
    - Select smallest subtree in parent 2 with same leaves
    - Exchange, prune duplicate leaves
  - Mutation: swap randomly selected leaves
  - Island model GA with migration



Testing GRAPHYL

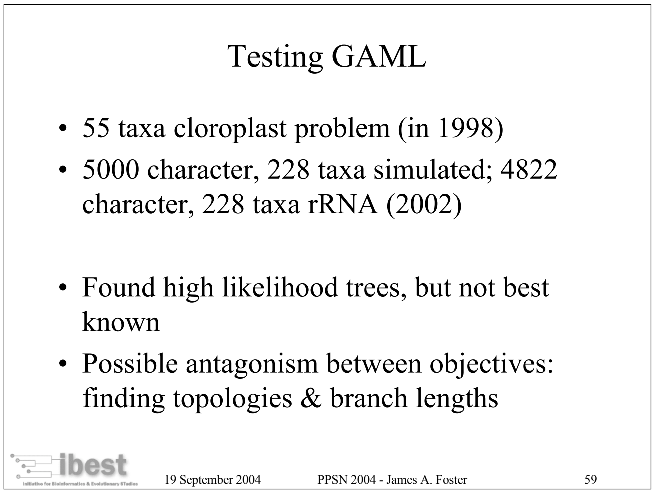
- Compare to Wagner parsimony utility in Phylip
  - on two datasets of binary characters: 23 species, 29 characters of Laminaflorii; 49 species, 61 attributes of angiosperm data
- GRAPHYL found comparable trees
  - Found <u>many</u> identical trees



GAML (Lewis)

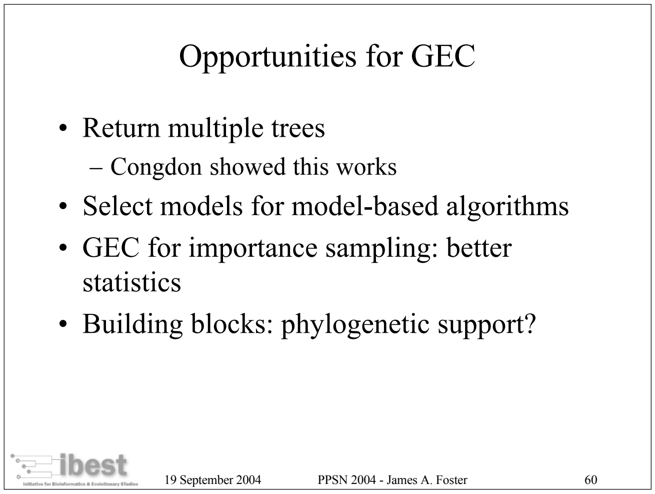
Find tree that maximizes (log)likelihood of data

- Representation: Tree with branch lengths, transition/transversion ratio (evolution model); Island model GA in 2002
- Fitness: ln(probability of sequences|Tree)
- Mutation
  - Random (with Gamma distribution) change in branch lengths
  - Random subtree pruning & Regrafting
  - Alter transition/transversion ratio
- Crossover (as in Evalyn)
  - Randomly select subtree in parent 1
  - Remove sequences in that subtree from Parent 2, simplify
  - Graft first subtree into Parent 2 at random point



Testing GAML

- 55 taxa cloroplast problem (in 1998)
- 5000 character, 228 taxa simulated; 4822 character, 228 taxa rRNA (2002)
- Found high likelihood trees, but not best known
- Possible antagonism between objectives: finding topologies & branch lengths



Opportunities for GEC

- Return multiple trees
  - Congdon showed this works
- Select models for model-based algorithms
- GEC for importance sampling: better statistics
- Building blocks: phylogenetic support?

Case Study
Gene Expression

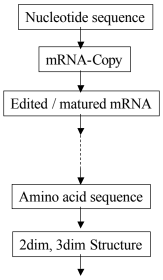
Mostly by Wolfgang Banzhaf

Outline

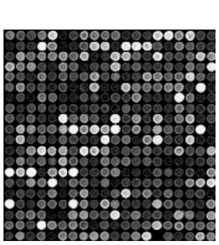
- What is the problem?
- What methods exist to solve it?
- What has been done with GEC?
- A specific approach
- Opportunities for GEC

What is the problem?

- Life is dynamic - DNA looks inert
- There needs to be a translation between the two
- Central dogma



Gene Expression Data



Fluorescence data regarding preferential hybridization of expressed genes (mRNA)

•Experimental noise (cross-hybridization, optical problems,…)
•Relative rather than absolute signals (background, etc)
•Different scanners, different chip manufacturers
• Artifacts from preparation of the cells (temperature dependence, concentrations of solvents, etc)
•Time-dependence of expression
•Large number of signals (features)

How can we discern patterns, e.g. of healthy vs. cancerous tissue?
Which are the genes that have most influence on the decision?

Aktivität in Reaktionen Verhalten Aktivität von I/O pairs

What is the problem? (II)

- Noise
- Experimental variation
- Abundance of features vs. small number of patterns (relative)
- High-dimensionality (absolute)

What methods exist to solve it?

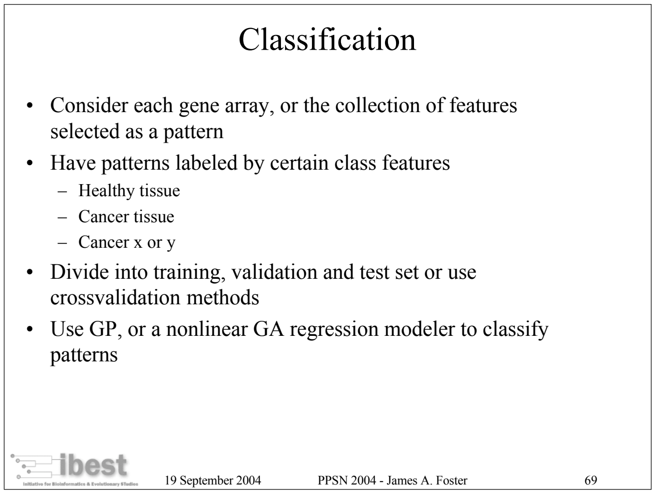
- Nearest neighbor
- Support vector machine
- Other machine learning approaches
- Self-organizing maps
- Other neuro and fuzzy approaches
- Evolutionary Computation

What EC is used for?

- Feature Selection
- Classification
- Discretization
- Other Parameter Optimization

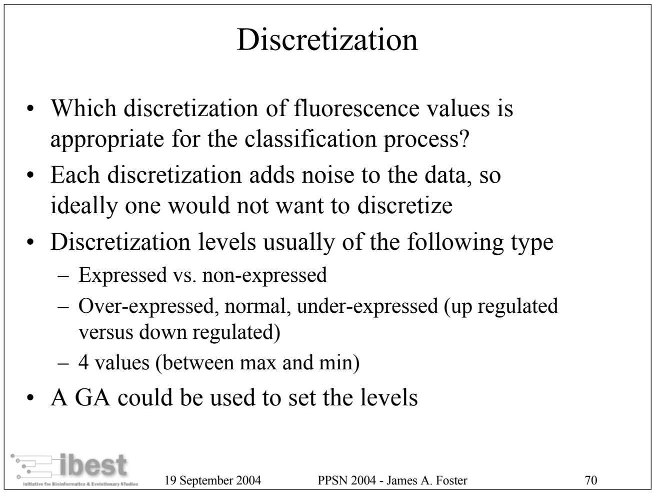
Feature Selection

- Which of the thousands of fluorescence spots should be used?
  - Take a selection
  - Take all
- What is a good feature selection method?
  - GP
  - GA



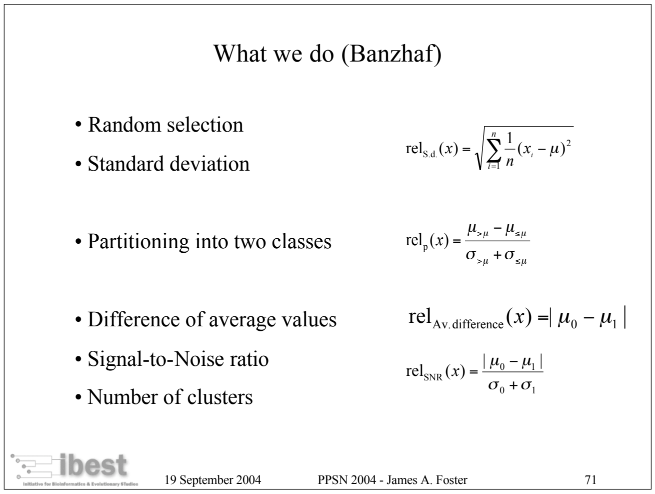
Classification

- Consider each gene array, or the collection of features selected as a pattern
- Have patterns labeled by certain class features
  - Healthy tissue
  - Cancer tissue
  - Cancer x or y
- Divide into training, validation and test set or use crossvalidation methods
- Use GP, or a nonlinear GA regression modeler to classify patterns



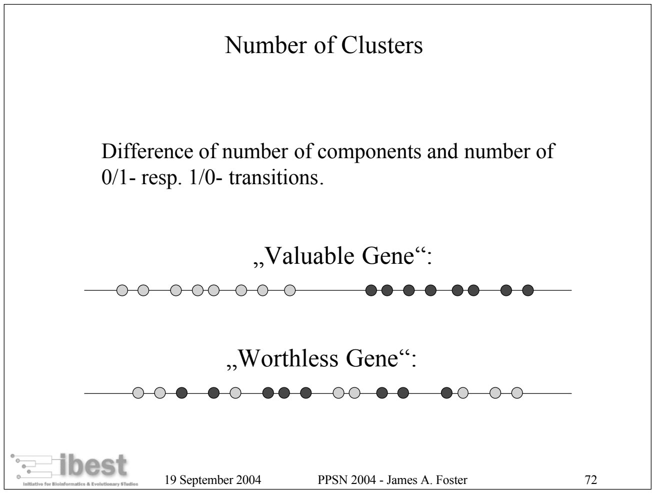
Discretization

- Which discretization of fluorescence values is appropriate for the classification process?
- Each discretization adds noise to the data, so ideally one would not want to discretize
- Discretization levels usually of the following type
  - Expressed vs. non-expressed
  - Over-expressed, normal, under-expressed (up regulated versus down regulated)
  - 4 values (between max and min)
- A GA could be used to set the levels



What we do (Banzhaf)

- Random selection
- Standard deviation $\mathrm{rel}_{\mathrm{S.d.}}(x) = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(x_i - \mu)^2}$
- Partitioning into two classes $\mathrm{rel}_{\mathrm{p}}(x) = \frac{\mu_{>\mu} - \mu_{\leq\mu}}{\sigma_{>\mu} + \sigma_{\leq\mu}}$
- Difference of average values $\mathrm{rel}_{\mathrm{Av.\,difference}}(x) = |\mu_0 - \mu_1|$
- Signal-to-Noise ratio $\mathrm{rel}_{\mathrm{SNR}}(x) = \frac{|\mu_0 - \mu_1|}{\sigma_0 + \sigma_1}$
- Number of clusters



Number of Clusters

Difference of number of components and number of 0/1- resp. 1/0- transitions.

„Valuable Gene“:

„Worthless Gene“:

GP-runs with Discipulus

- Binary Classification
- 3 subsets: Training, Validation, Applied
- Standard parameters
- 100 runs per experiment
- Different numbers of features

Results (10 Features)

Selection	**Colon Tumor**	**ALL/AML**
Standard Deviation	100 / 86 / 75	100 / 75 / 67
Two Partition	**100 / 90 / 95**	**100 / 100 / 100**
Mean Difference	100 / 90 / 75	100 / 96 / 96
Signal-to-Noise	100 / 100 / 80	100 / 100 / 96
Cluster Count	100 / 95 / 80	100 / 100 / 96
Random (Mean values)	96 / 90 / 71	97 / 85 / 72

Other Methods

Selection	**Colon Tumor**	**ALL/AML**
GA	/ / 55 (Li+,2001)	-
SVM	/ / 90 (Furey+,2000)	-
Neighborhood anal.	-	/ / 85 (Golub+, 1999)
Selective Expression	-	/ / 100 (Aris+, 2002)
Double conjugated Clustering	-	/ / 100 (Busygin+, 2002)

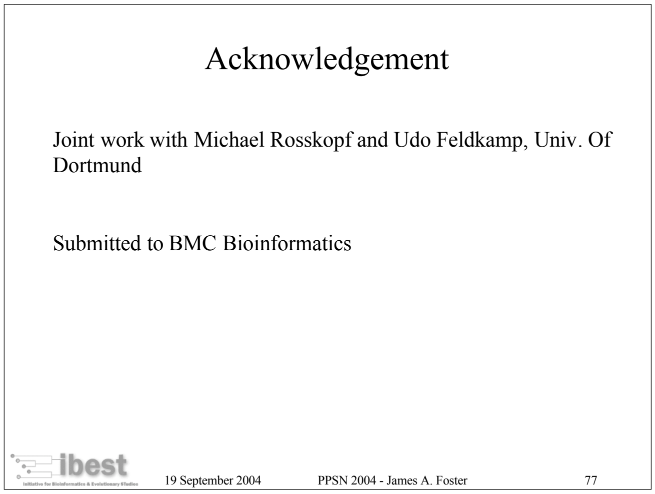
Summary

Main task of static gene array data evaluation is pattern recognition and classification

Generally, EC methods have been shown to be very competitive on such tasks

Feature selection particularly is a strength of GA, and notably GP approaches (in contrast to NN, for instance)

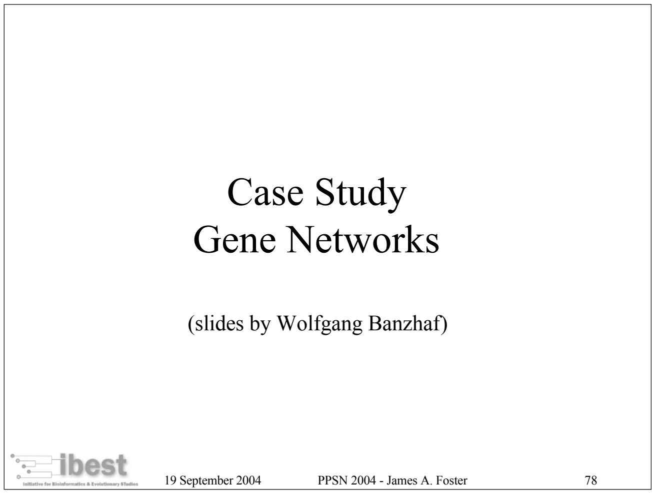
Multi-class prediction is generally more difficult than two class problems



Acknowledgement

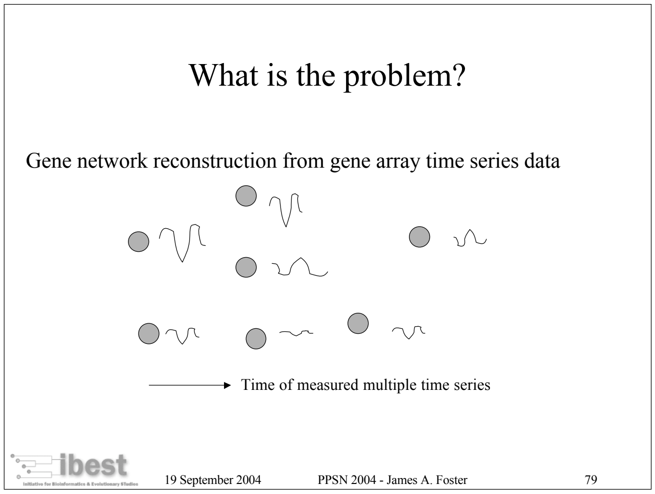
Joint work with Michael Rosskopf and Udo Feldkamp, Univ. Of Dortmund

Submitted to BMC Bioinformatics



Case Study
Gene Networks

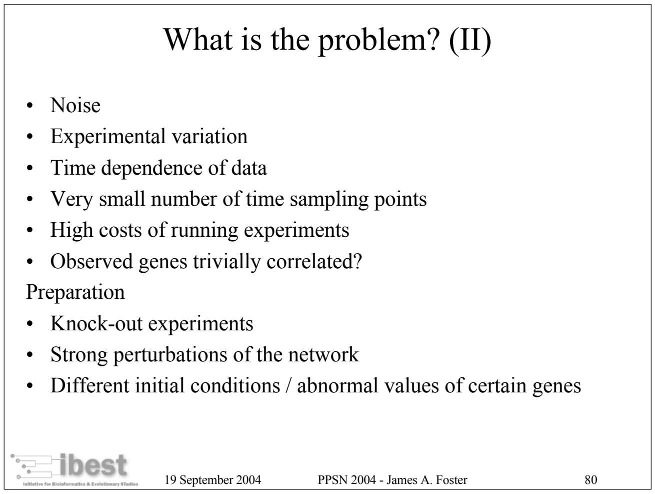
(slides by Wolfgang Banzhaf)



What is the problem?

Gene network reconstruction from gene array time series data

Time of measured multiple time series



What is the problem? (II)

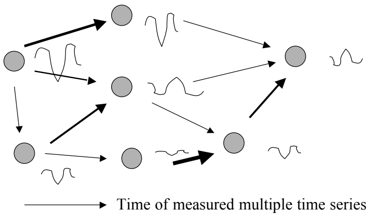
- Noise
- Experimental variation
- Time dependence of data
- Very small number of time sampling points
- High costs of running experiments
- Observed genes trivially correlated?

Preparation

- Knock-out experiments
- Strong perturbations of the network
- Different initial conditions / abnormal values of certain genes

The goal

Goal of network reconstruction: Determine the links between genes



The more data the more restrictions for links, the more unique the solution

What methods exist to solve it?

- Boolean networks (binary values)
- Discrete networks a la R. Thomas
- Weight Matrices
- Bayesian Networks (no cycles)
- Dynamic Bayesian Networks
- Differential equations (small numbers of genes)
- Difference Equations
- EC techniques

- Researchers generally prefer statistical methods due to the noise inherent in the experimental techniques
- Large number of possibilities: 20 genes -> 10^72 DAGs

What has been done with GEC?

- Parameter optimization for differential equation coefficients
- Bayesian Networks (node ordering)
- Belief Network coefficients
- Boolean Network coefficients
- Time series prediction using GA/ES/GP

A specific approach

Ando/Iba: Construction of a Genetic Network using an Evolutionary Algorithm and Combined Fitness Function (Genome Informatics 14, 2003, pp.94-103)

Parameter Estimation of a S-system network model: Excitatory and inhibitory (nonlinear) regulation of genes. Parameters are $(\alpha_i, \beta_i, g_{ij}, h_{ij})$

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{N} x_j^{g_{ij}} - \prod_{j=1}^{N} x_j^{h_{ij}}$$

Noise model between true expression $S(x_k)$ and measured/assumed expression x_k is Gaussian with std.dev constant over time and equal over all genes

$$\varepsilon_k = N(x_k - S(x_k), \sigma)$$

Fitness function to minimize: AIC=log-likelihood of a model + # degrees of freedom

$$\Lambda(M, \sigma) = -\frac{1}{2\sigma^2} \sum_{t=1}^{T} [x(t) - S(x(t)]^2 - \frac{T}{2} \ln(2\pi\sigma^2)$$

A specific approach II

2 phase GA: (1) Estimation of parameters for each gene and its regulators $(\alpha_i, \beta_i, g_{ij}, h_{ij})$
(2) Estimation of parameters of the whole network

Tests with artificial data (networks artificially generated + noise added)

E.Coli data: Tryptophan metabolism (PNAS 97 (2000) 12170-12175)
3 time series, 5 time points each (starvation and overdose of tryptophan)

Results: Artificial networks well reconstructed, simple natural system with problems due to noise levels.

Opportunities for GEC

- General ODE models for gene expression (eg by GP)
- Devise good fitness functions
- Learn to deal with noise levels
- Study artificial models of regulatory networks (what is optimized by a regulatory network?)
- Include prior knowledge into modeling tool
- Suggest a set of models instead of just one

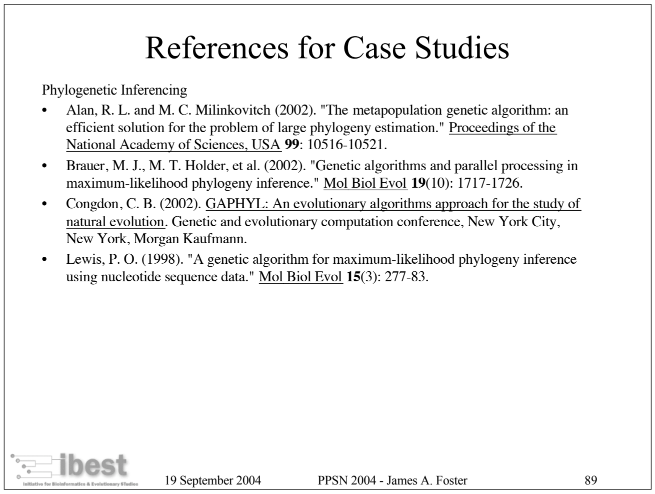
Further information

References for Case Studies

This is NOT a complete bibliography, it only lists work discussed directly in this tutorial

Multiple Sequence Alignment

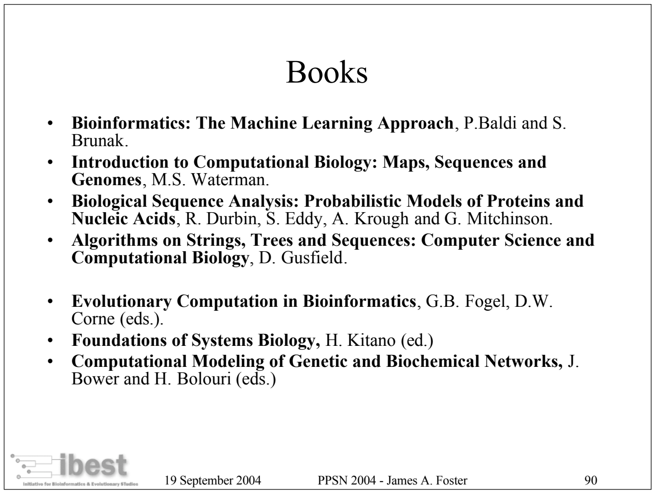
- Feng, D. F. and R. F. Doolittle (1996). "Progressive alignment of amino acid sequences and construction of phylogenetic trees from them." Methods Enzymol **266**: 368-82..
- Notredame, C. and D. G. Higgins (1996). "SAGA: sequence alignment by genetic algorithm." Nucleic Acids Research **24**(8): 1515-1524.
- Shyu, C. and J. A. Foster (2003). Evolving consensus sequence for multiple sequence alignment with a genetic algorithm. Proc. Genetic and Evolutionary Computing Conference (GECCO), Chicago, Springer-Verlag.
- Shyu, C., L. Sheneman, et al. (in press). "Evolutionary computation for multiple sequence alignment." Genetic Programming and Evolvable Machines.
- Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.
- Thompson, J. D., F. Plewniak, et al. (1999). "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs." Bioinformatics **15**(1): 97-98.
- Wang, L. and T. Jiang (1994). "On the complexity of multiple sequence alignment." J Comput Biol **1**(4): 337-48.
- Zhang, C. and A. K. Wong (1997). "A genetic algorithm for multiple molecular sequence alignment." Comput Appl Biosci **13**(6): 565-81.



References for Case Studies

Phylogenetic Inferencing

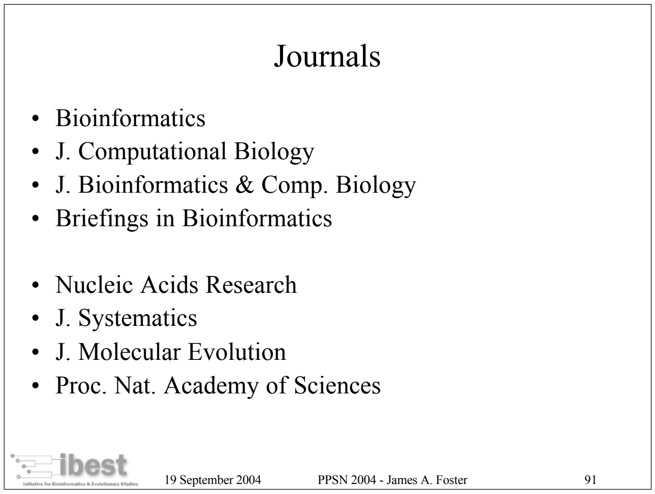
- Alan, R. L. and M. C. Milinkovitch (2002). "The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation." Proceedings of the National Academy of Sciences, USA **99**: 10516-10521.
- Brauer, M. J., M. T. Holder, et al. (2002). "Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference." Mol Biol Evol **19**(10): 1717-1726.
- Congdon, C. B. (2002). GAPHYL: An evolutionary algorithms approach for the study of natural evolution. Genetic and evolutionary computation conference, New York City, New York, Morgan Kaufmann.
- Lewis, P. O. (1998). "A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data." Mol Biol Evol **15**(3): 277-83.



Books

- **Bioinformatics: The Machine Learning Approach**, P.Baldi and S. Brunak.
- **Introduction to Computational Biology: Maps, Sequences and Genomes**, M.S. Waterman.
- **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**, R. Durbin, S. Eddy, A. Krough and G. Mitchinson.
- **Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology**, D. Gusfield.

- **Evolutionary Computation in Bioinformatics**, G.B. Fogel, D.W. Corne (eds.).
- **Foundations of Systems Biology,** H. Kitano (ed.)
- **Computational Modeling of Genetic and Biochemical Networks,** J. Bower and H. Bolouri (eds.)



Journals

- Bioinformatics
- J. Computational Biology
- J. Bioinformatics & Comp. Biology
- Briefings in Bioinformatics

- Nucleic Acids Research
- J. Systematics
- J. Molecular Evolution
- Proc. Nat. Academy of Sciences

Webpages

- Int. Soc. For Comp. Bio: www.iscb.org
- Tutorials: www.techfak.uni-bielefeld.de/bcd/original-welcome.html
- NIH/NCBI: www.ncbi.nlm.nih.gov/Education/index.html
- Bioplanet: www.bioplanet.com/links.htm

Conferences

- PPSN, of course!
- Pacific Symposium on Biocomputing (PSB)
- Research in Computational Biology (Recomb)
- Intelligent Systems in Molecular Biology (ISMB)
- Genetic & Evolutionary Computation Conference (GECCO)
- Congress on Evolutionary Computation (CEC)
- Evolution meetings: evolution04.biology.colostate.edu

Graduate school

- U. Idaho Bioinformatics and Computational Biology (MS/PhD)
- Memorial U Computational Science Program (MS)
- List: www.iscb.org/univ_programs/program_board.php

(almost) Final thoughts

GEC may be useful for several biological problems:

- Predicting alternate splice sites
- Mapping exon groups to proteins
- Predicting RNAi and other RNA editing
- Protein folding
- Predicting structure of proteins or RNA
- Modeling interacting molecules or organisms
- Inferring expression/metabolic/developmental networks
- Mining existing data
- Correlating disease and other data
- And much much more…

(Nearly) final thoughts

- Challenges
  - How to test GECs convincingly?
    - Simulation data: not “biological”
    - Natural data: poorly understood
    - Benchmark databases: many hidden pitfalls
  - Biological data is
    - Too sparse or too rich
    - Expensive to acquire
    - Full of errors and ambiguities
    - Derived from molecules or critters (not maths)

Final thoughts

- Suggestions
  - Get to know some biologists
  - Be patient, brave, honest
  - Work on *their* problems
  - Don’t be satisfied until *they* are
  - Publish in *their* journals, present posters in *their* conferences

No thoughts are truly “final”