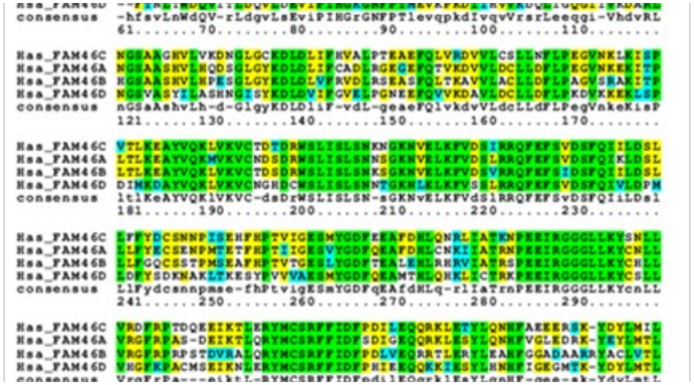
Algorithmic Foundations of Computational Biology
CSCI 1820/2820
An overview

- **Ch. 1 The BLAST Algorithm and Karlin-Altschul Statistics**
- **Ch. 2 Genome Assembly Algorithms and Haplotype Assembly Algorithms**
- **Ch. 3 The Protein Folding Problem: From HP-lattice models to AlphaFold**
- **Ch. 4 Recombination and Ancestral Recombination Graphs (ARGs) Algorithms**
- **Ch.5 Rigorous clustering: Spectral Graph Theory Algorithms**
- **Ch. 6 The Regulatory Genome and Gene Regulatory Networks**

Ch. 1: BLAST Algorithm



```
consensus   -hfsvLnWdQV-rLdgvLsEviPIHGrGNFPTlevqpkdIvqvVrsrLeeqgi-VhdvRL
            61.......70........80........90........100.......110........

Has_FAM46C  NGSAAGHVLVKDNGLGCKDLDLIFHVALPTEAEFQLVRDVVLCSLLNFLPEGVNKLKISP
Hsa_FAM46A  NGSAASHVLHQDSGLGYKDLDLIFCADLRGEGEFQTVKDVVLDCLLDFLPEGVNKEKITP
Hsa_FAM46B  HGSAASHVLHPESGLGYKDLDLVFRVDLRSEASFQLTKAVVLACLLDFLPAGVSRAKITP
Hsa_FAM46D  NGSVASYILASHNGISYKDLDVIFGVELPGNEEFQVVKDAVLDCLLDFLPKDVKKEKLSP
consensus   nGSaAshvLh-d-GlgyKDLDliF-vdL-geaeFQlvkdvVLdcLLdFLPegVnkeKisP
            121......130.......140.......150.......160.......170........

Has_FAM46C  VTLKEAYVQKLVKVCTDTDRWSLISLSNKNGKNVELKFVDSIRRQFEFSVDSFQIILDSL
Hsa_FAM46A  LTLKEAYVQKMVKVCNDSDRWSLISLSNNSGKNVELKFVDSLRRQFEFSVDSFQIKLDSL
Hsa_FAM46B  LTLKEAYVQKLVKVCTDSDRWSLISLSNKSGKNVELKFVDSVRRQFEFSIDSFQIILDSL
Hsa_FAM46D  DIMKDAYVQKLVKVCNGHDCWSLISLSNNTGKNLELKFVSSLRRQFEFSVDSFQIVLDPM
consensus   ltlKeAYVQKlVKVC-dsDrWSLISLSN-sGKNvELKFVdSlRRQFEFSvDSFQIiLDsl
            181......190.......200.......210.......220.......230........

Has_FAM46C  LFFYDCSNNPISEHFHPTVIGESMYGDFEEAFDHLQNRLIATKNPEEIRGGGLLKYSNLL
Hsa_FAM46A  LLFYECSENPMTETFHPTIIGESVYGDFQEAFDHLCNKIIATRNPEEIRGGGLLKYCNLL
Hsa_FAM46B  LLFGQCSSTPMSEAFHPTVTGESLYGDFTEALEHLRHRVIATRSPEEIRGGGLLKYCHLL
Hsa_FAM46D  LDFYSDKNAKLTKESYPVVVAESMYGDFQEAMTHLQHKLICTRKPEEIRGGGLLKYCSLL
consensus   LlFydcsnnpmse-fhPtvigESmYGDFqEAfdHLq-rlIaTrnPEEIRGGGLLKYcnLL
            241......250.......260.......270.......280.......290........

Has_FAM46C  VRDFRPTDQEEIKTLERYMCSRFFIDFPDILEQQRKLETYLQNHFAEEERSK-YDYLMIL
Hsa_FAM46A  VRGFRPAS-DEIKTLQRYMCSRFFIDFSDIGEQQRKLESYLQNHFVGLEDRK-YEYLMTL
Hsa_FAM46B  VRGFRPRPSTDVRALQRYMCSRFFIDFPDLVEQRRTLERYLEAHFGGADAARRYACLVTL
Hsa_FAM46D  VHGFKPACMSEIKNLERYMCSRFFIDFPHIEEQQKKIESYLHNHFIGEGMTK-YDYLMTL
consensus   VrgFrPa---eiktL-RYMCSRFFIDFpdilEQqrklEsYLqnHF-gee-ak-YdvLmtL
```

Given a biomolecular query sequence Q
and a database DB of biomolecular sequences

Find all the biomolecular sequences in DB that have high alignment scores to the query

Biomolecular: DNA, RNA, protein

Problems we need to solve along the way:

Problem 1. General scoring schemes as hypotheses testing frameworks
The Karlin-Altschul Statistics and the max scoring subsequence

Problem 2. Random Walks Theory and The Gambler's Ruin's Problem

Problem 3. De-noising: how long an alignment needs to be non-random?

Problem 4. Information Theory and the theory of scoring matrices for alignment

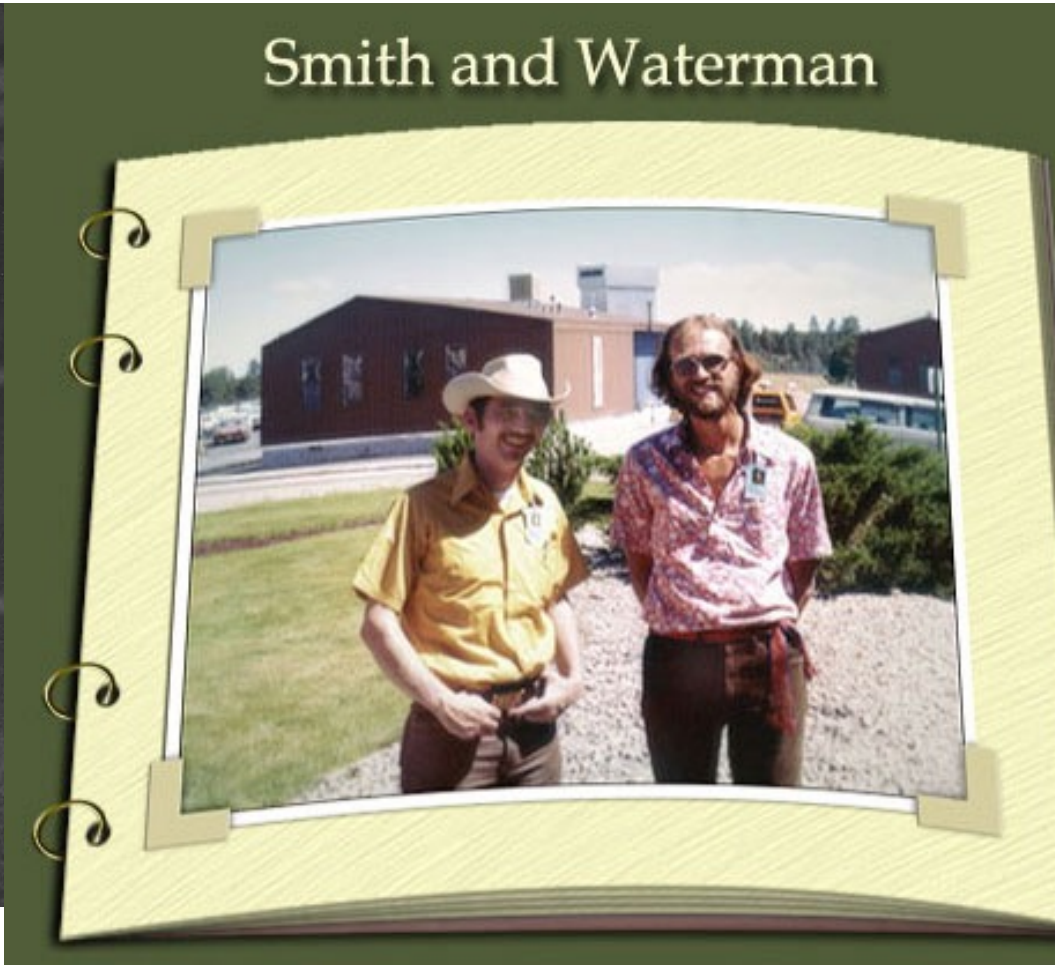
Dr. Margaret Oakley Dayhoff

The Mother & Father of Bioinformatics

Temple Smith and Michael Waterman

at Los Alamos, New Mexico

Photo by David Lipman, Taken Summer of 1980



Sir Ronald Aylmer Fisher

The Lady Tasting Tea Problem

the

Null Hypothesis

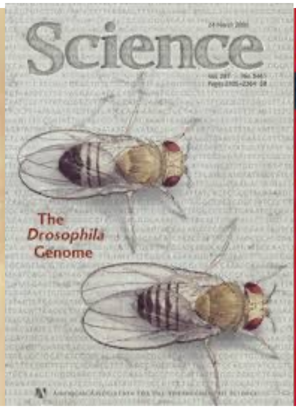
Born	Ronald Aylmer Fisher 17 February 1890 London, England
Died	29 July 1962 (aged 72) Adelaide, South Australia, Australia

•Linear discriminant analysis is a generalization of Fisher's li discriminant[47][83]
•Fisher information, see also scoring algorithm also known a scoring, and Minimum Fisher information, a variational prin which, when applied with the proper constraints needed to empirically known expectation values, determines the best distribution that characterizes the system.[84]
•*F*-distribution, arises frequently as the null distribution of a statistic, most notably in the analysis of variance
•Fisher–Tippett–Gnedenko theorem : Fisher's contribution t made in 1927
•Fisher–Tippett distribution
•Fisher-Yates shuffle algorithm
•Von Mises–Fisher distribution[85]
•Inverse probability, a term Fisher used in 1922, referring to fundamental paradox of inverse probability" as the source o confusion between statistical terms which refer to the true estimated, with the actual value arrived at by estimation, w subject to error.[86]
•Fisher's permutation test
•Fisher's inequality[87]
•Sufficient statistic, when a statistic is *sufficient* with respect a statistical model and its associated unknown parameter if statistic that can be calculated from the same sample provic additional information as to the value of the parameter".[88]
•Fisher's noncentral hypergeometric distribution, a generali the hypergeometric distribution, where sampling probabiliti modified by weight factors.
•Student's *t*-distribution, widely used in statistics.[89][90]
•The concept of an ancillary statistic and the notion (the and principle) that one should condition on ancillary statistics.

The BLAST algorithm

- Detect all *word hits* (exact, or nearly identical matches) of a given length between the two sequences
  - k=10 for nucleotide sequences (exact word matches)
  - k=3 for protein sequences (nearly identical word matches)
- Extend the word hits in both directions to high-scoring *gap-free* segment pairs (HSPs)
  - retain only HSPs that score above a threshold
  - start from the center of the HSP (original BLAST, 1990), or from the center of a pair of HSPs located close to each other on the same diagonal (gapped BLAST, 1997)
- Extend the HSPs in both directions allowing for gaps
  - use dynamic programming, and stop when the alignment score falls more than a threshold X below the best score yet seen
- Report all statistically significant local alignments
  - E-value (starting with BLAST 2.0) is used to measure the statistical significance
  - *E-value* = the number of alignments with score equal to or higher than *s* one would expect to find by chance when searching the database

Ch. 2: Genome Assembly Algorithms



Questions: What algorithms to use to assemble DNA pieces into contigs and scafolds?
How long are the contigs?
How much the DNA target region is covered by the contigs?
How to measure the success of a genome assembly?

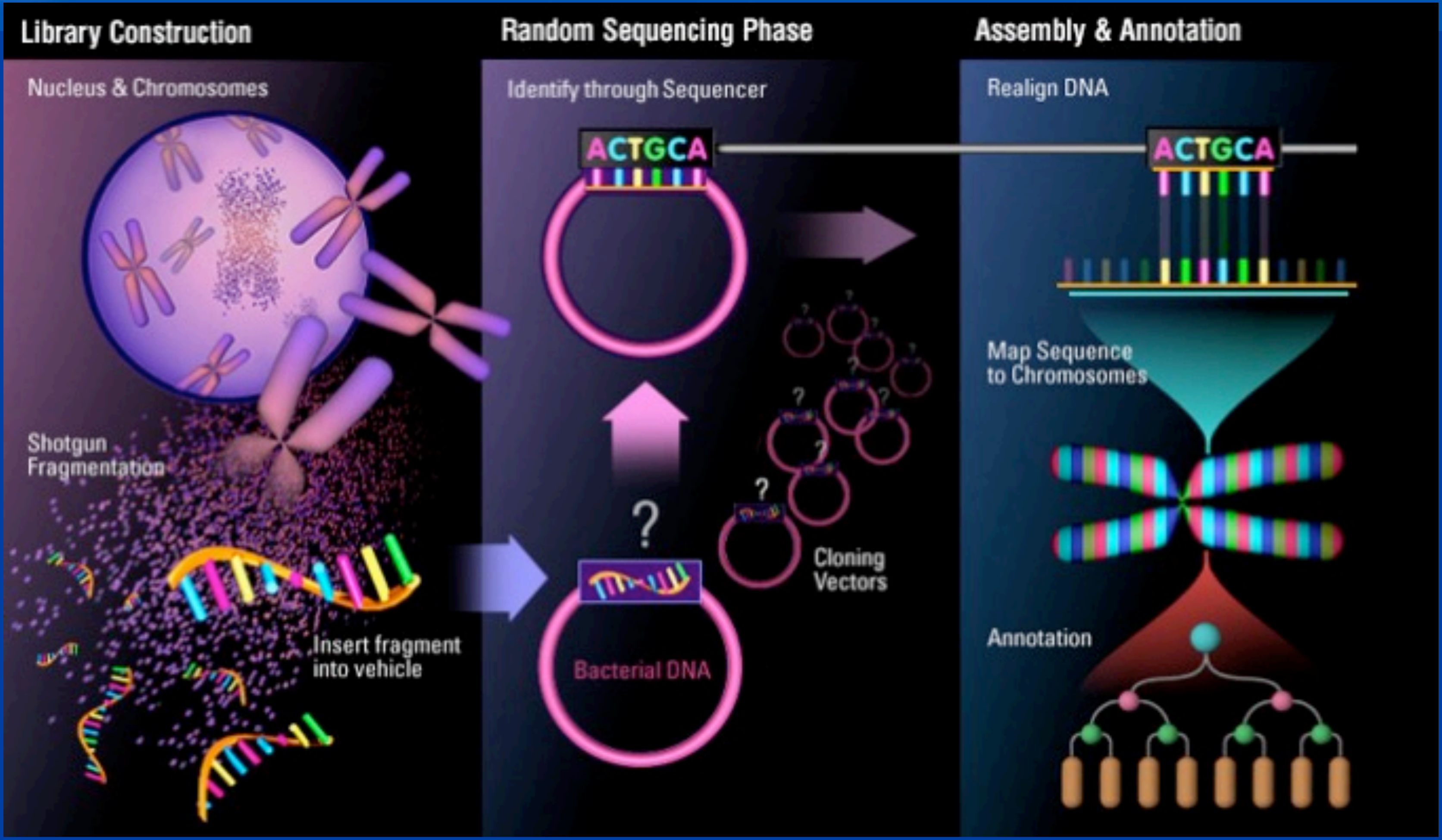
Problems we need to solve along the way

Problem 1. Genome Assembly Algorithms

Problem 2. Poisson statistics for DNA and Genome Assembly

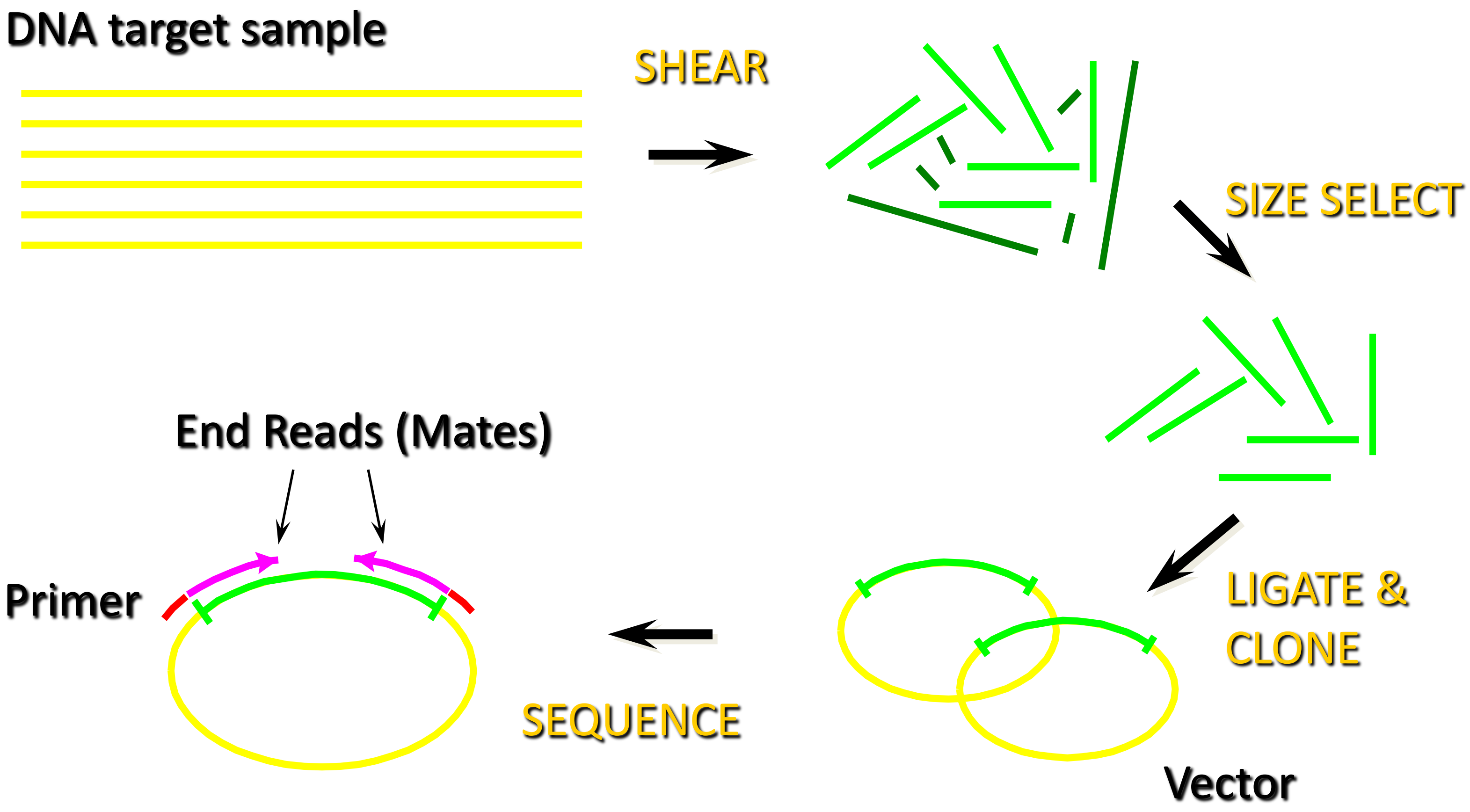
Problem 3. Ham Smith's DNA Lab with no windows

Whole Genome Shotgun Sequencing

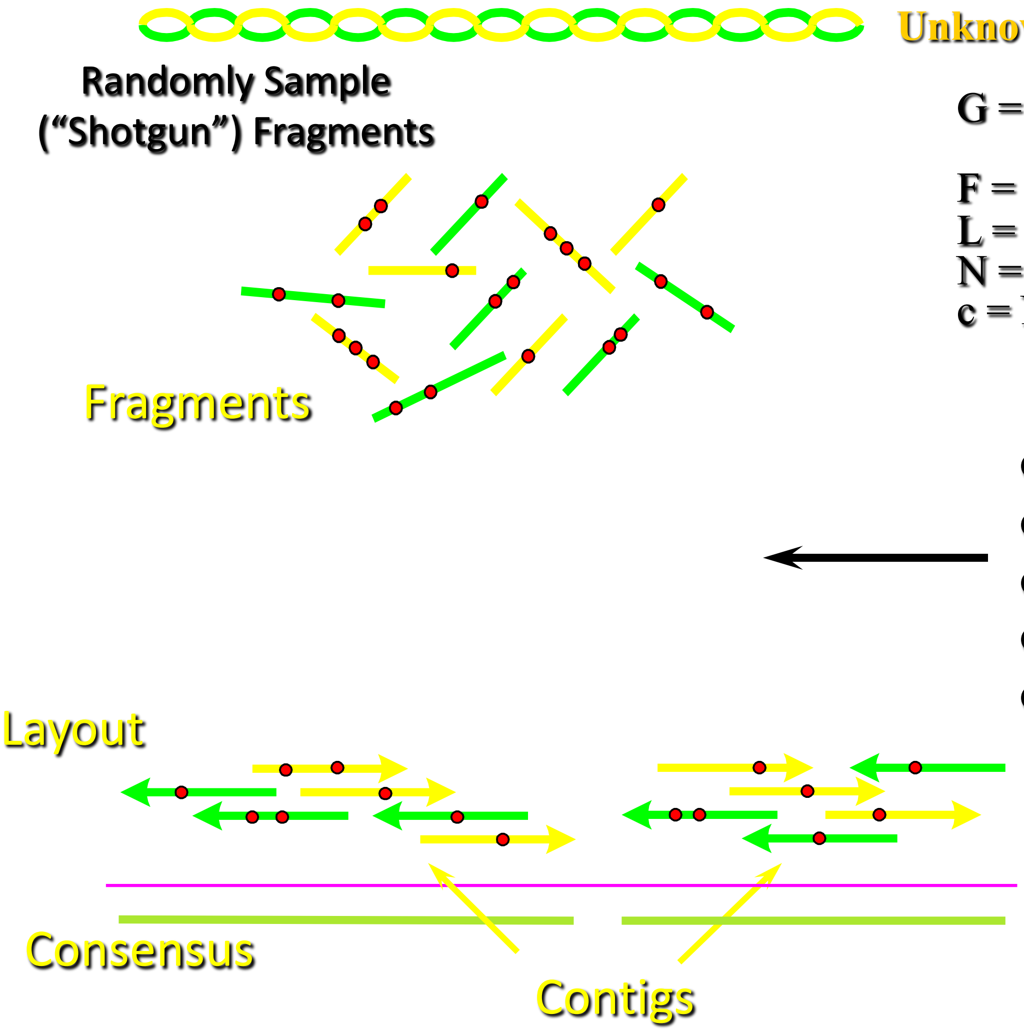


CELERA
A PE Corporation Business

Shotgun DNA Sequencing (Technology)



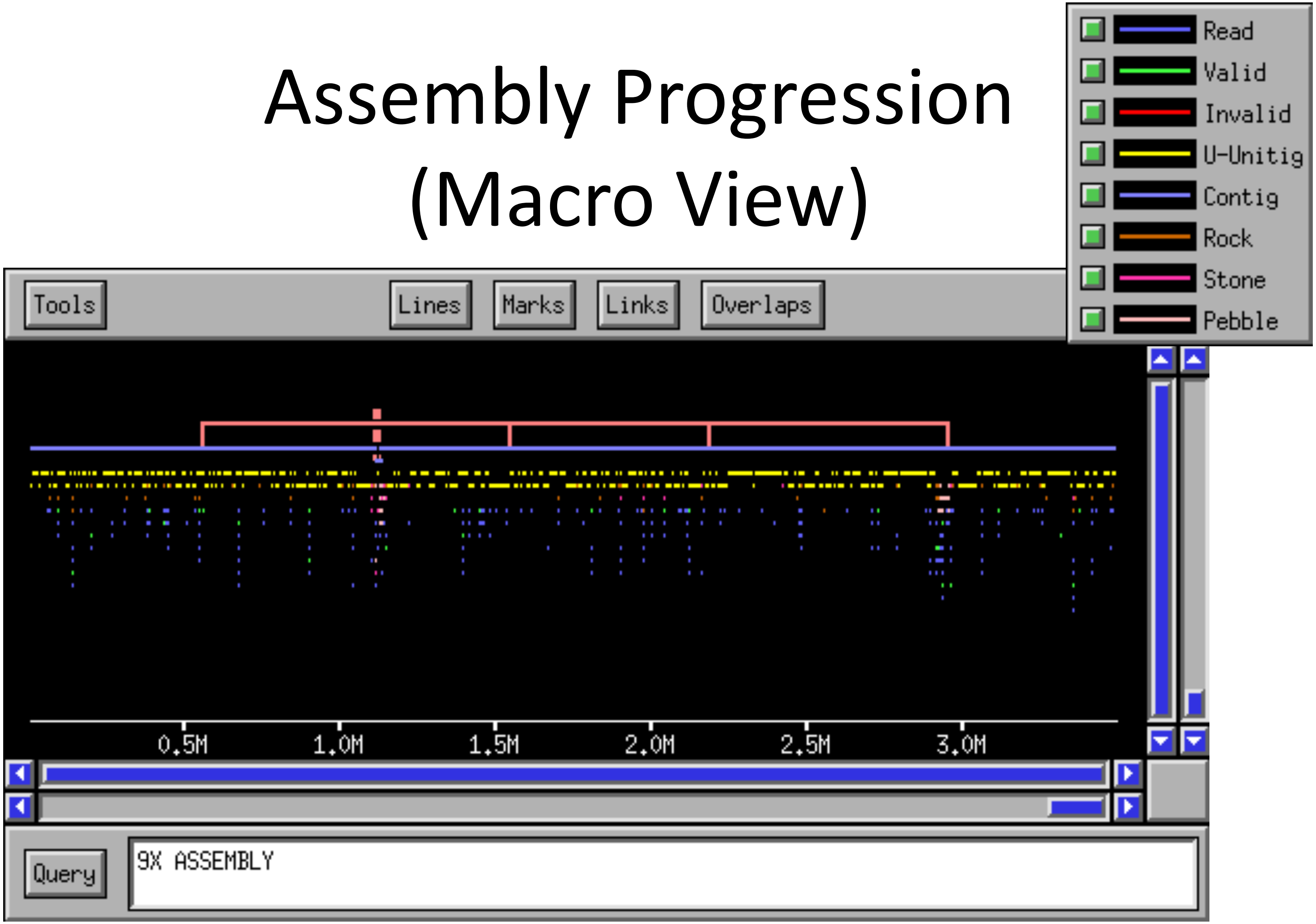
Shotgun DNA Sequencing (Computation)



G = 100Kbp	Target Length (e.g., BAC, P1, PAC)
F = 1600	# of Fragments
L = 500	Avg. Fragment Length
N = FL = 800Kbp	Total Bases Sequenced
c = N/G = 8	Avg. Coverage

- ● UNKNOWN ORIENTATION
- ● SEQUENCING ERRORS
- ● INCOMPLETE COVERAGE
- ● CONSTRAINTS (MATES)
- ● REPEATS

Assembly Progression (Macro View)

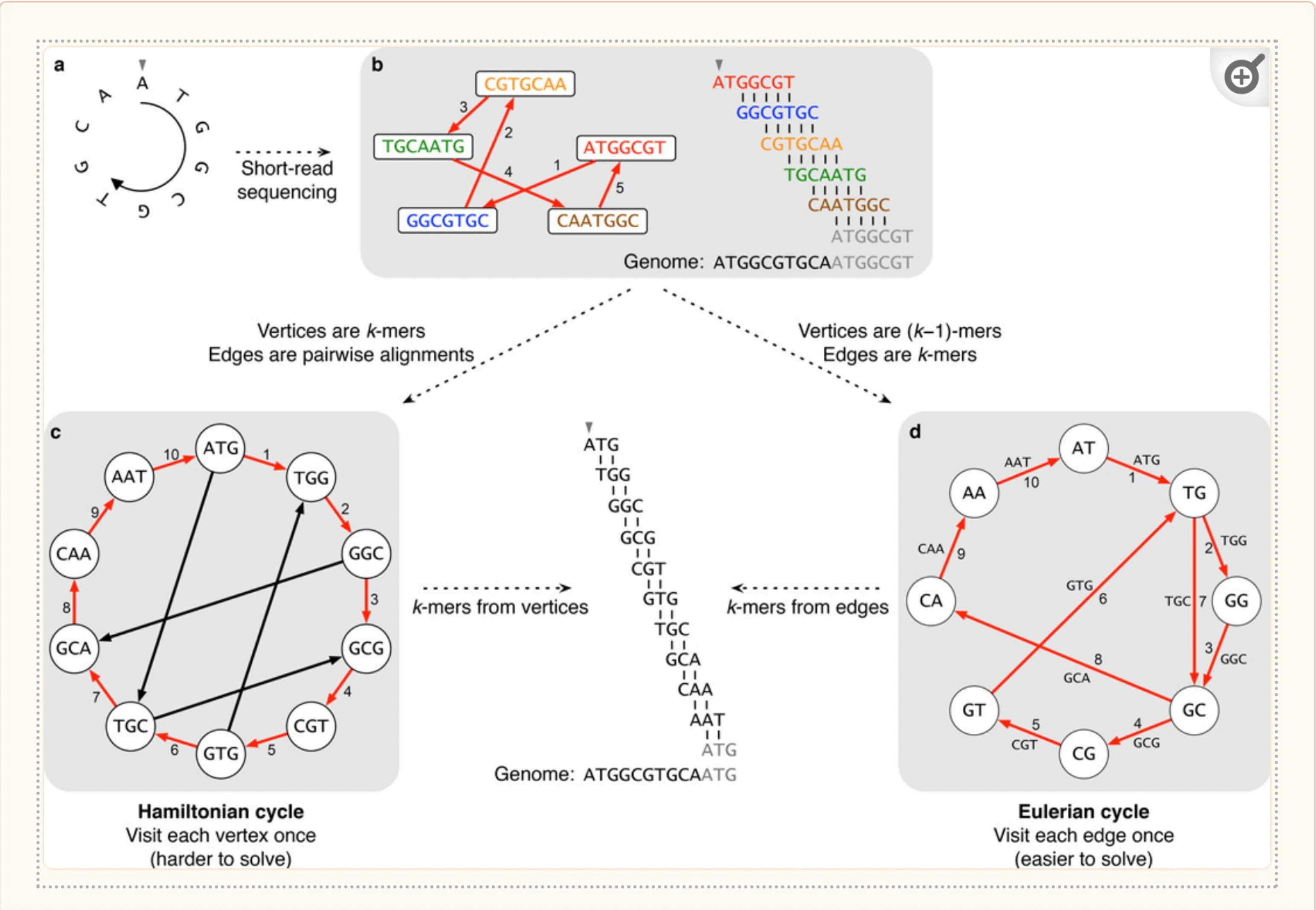


9X ASSEMBLY

Siméon Denis Poisson

Born	21 June 1781 Pithiviers, Kingdom of France (present-day Loiret)
Died	25 April 1840 (aged 58) Sceaux, Hauts-de-Seine, Kingdom of France
Alma mater	École Polytechnique
Known for	Poisson process Poisson equation Poisson kernel Poisson distribution Poisson limit theorem Poisson bracket Poisson algebra Poisson regression Poisson summation formula Poisson's spot Poisson's ratio Poisson zeros Conway–Maxwell–Poisson distribution Euler–Poisson–Darboux equation
Scientific career	
Fields	Mathematics and physics
Institutions	École Polytechnique Bureau des Longitudes Faculté des sciences de Paris École de Saint-Cyr
Academic advisors	Joseph-Louis Lagrange Pierre-Simon Laplace
Doctoral students	Michel Chasles Joseph Liouville
Other notable students	Nicolas Léonard Sadi Carnot Peter Gustav Lejeune Dirichlet

de Bruijn Genome Assembly



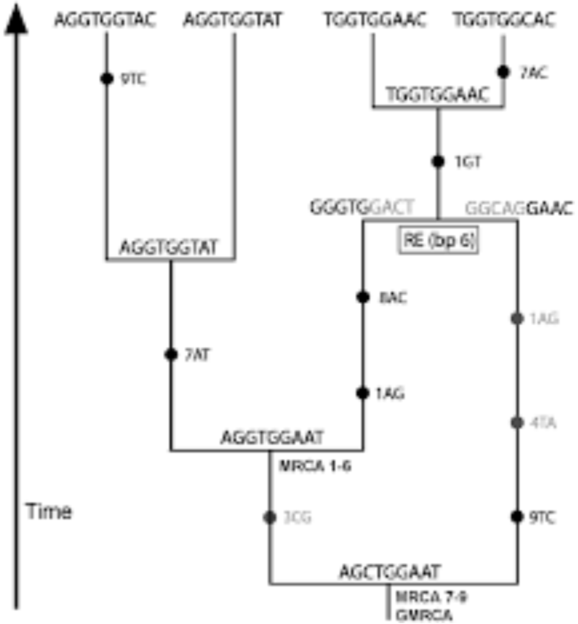
Vertices are *k*-mers
Edges are pairwise alignments

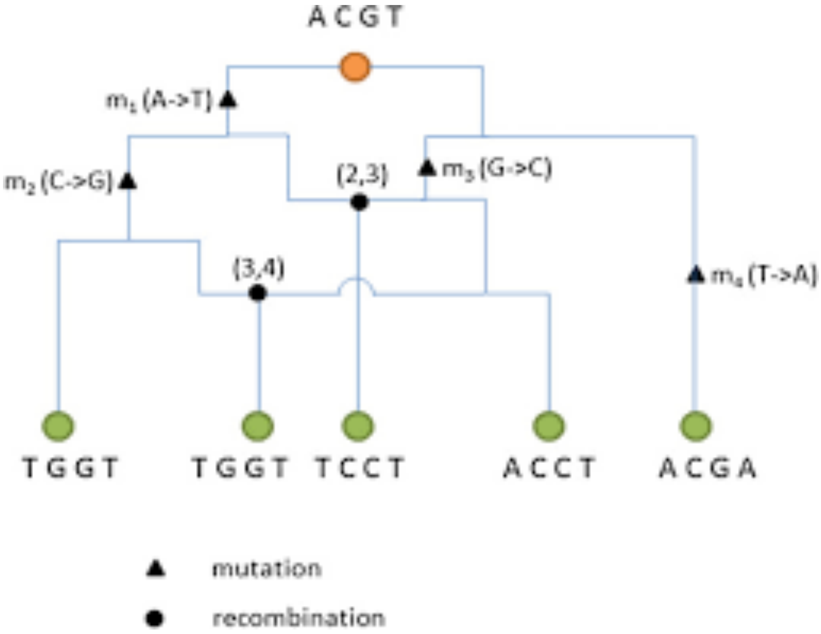
Vertices are (*k*–1)-mers
Edges are *k*-mers

Hamiltonian cycle
Visit each vertex once
(harder to solve)

Eulerian cycle
Visit each edge once
(easier to solve)

Ch. 3 Recombination and Ancestral Recombination Graphs (ARG) Algorithms

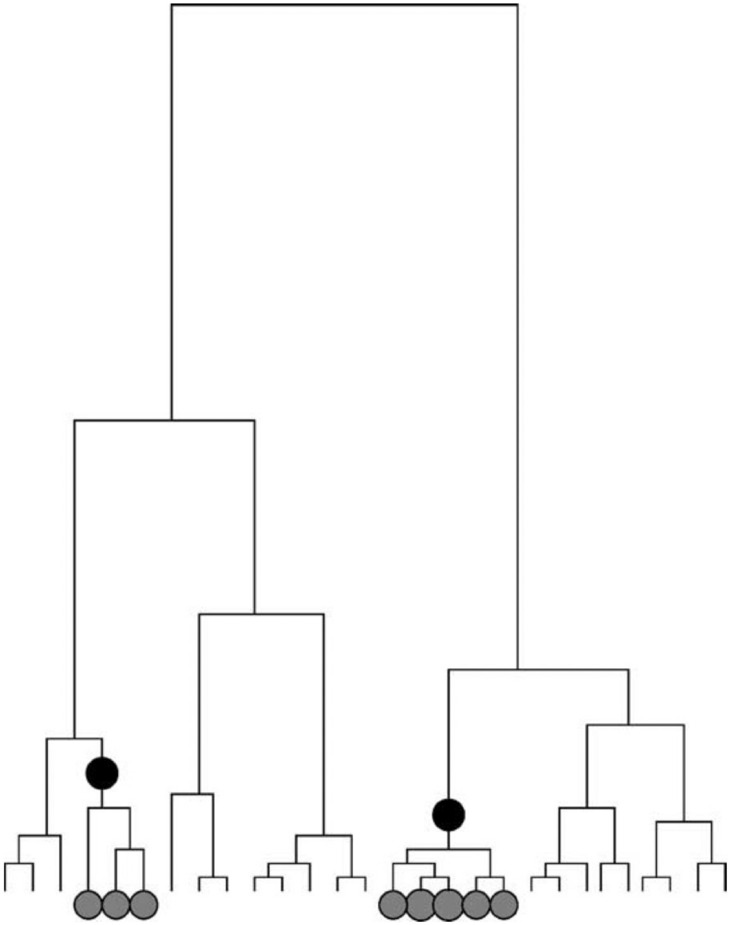


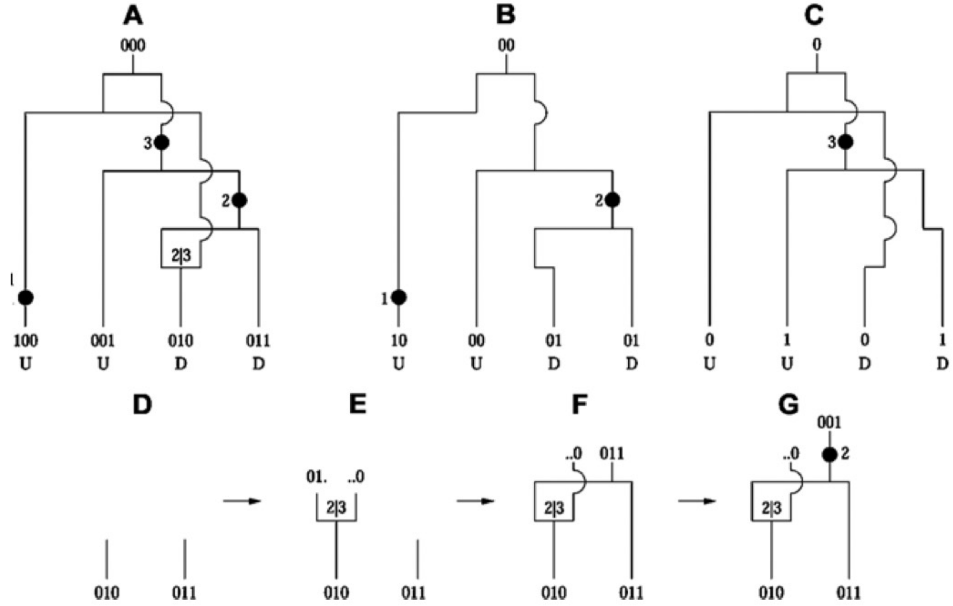


How do we reconstruct genealogies of a sample of individuals incorporating past mutations and recombinations?

Recombination + Phylogenetic Trees = ARG

Ancestral Recombination Graph and Marginal Trees





Ch. 5 Rigorous Clustering Algorithms Spectral Graph Theory Algorithms

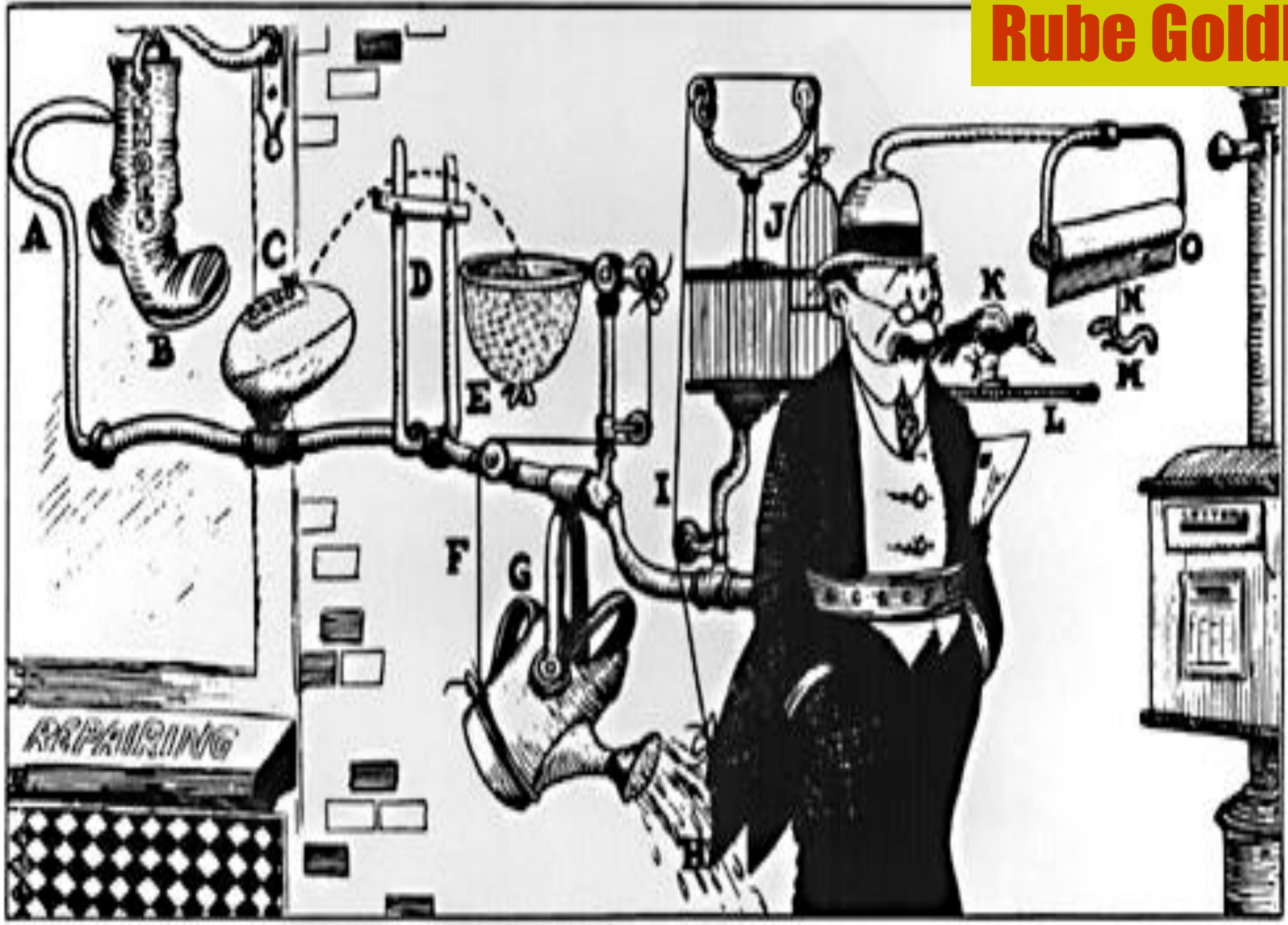
Algorithms and Statistical Theory

- An introduction to Linear Algebra foundations for graph theory
- Principles of Clustering Theory
- Graph Laplacians
- Graph cuts and random walks intuitions for Spectral Clustering
- Unnormalized Spectral Clustering Algorithms
- Normalized Spectral Clustering Algorithms
- Algorithmic Fairness and Clustering

Pierre-Simon Laplace

Born	23 March 1749 Beaumont-en-Auge, Normandy, Kingdom of France
Died	5 March 1827 (aged 77) Paris, Kingdom of France
Alma mater	University of Caen
Known for	show
Scientific career	
Fields	Astronomy and Mathematics
Notable students	Siméon Denis Poisson Napoleon Bonaparte

Rube Goldberg's Innovation



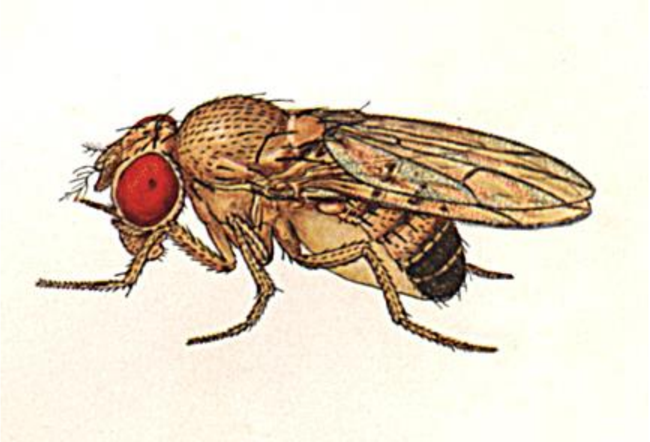
Keep You From Forgetting To Mail Your Wife's Letter RUBE GOLDBERG (tm) RGI 049

Mixed character of the problem :

continuous mathematics
discrete mathematics

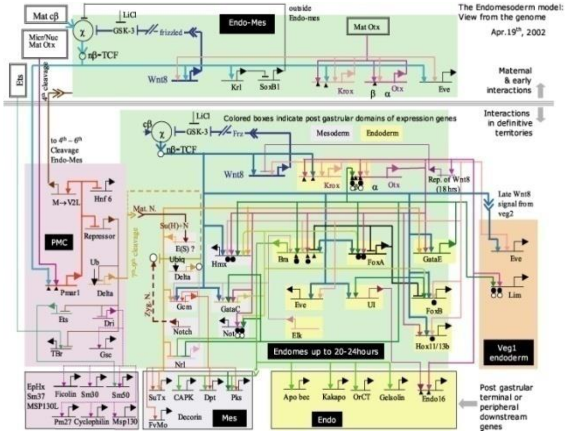
GENOMIC REGULATORY SYSTEMS

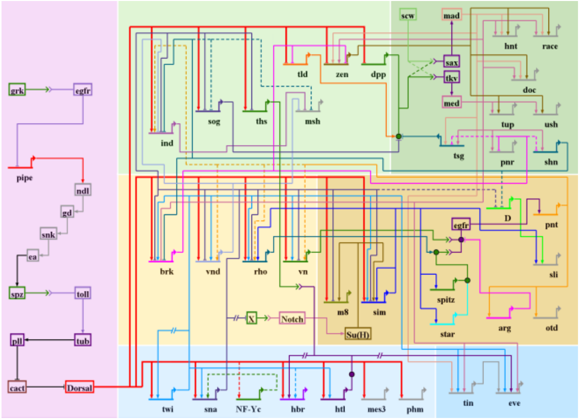
A Tale of Two Networks



Sea Urchin

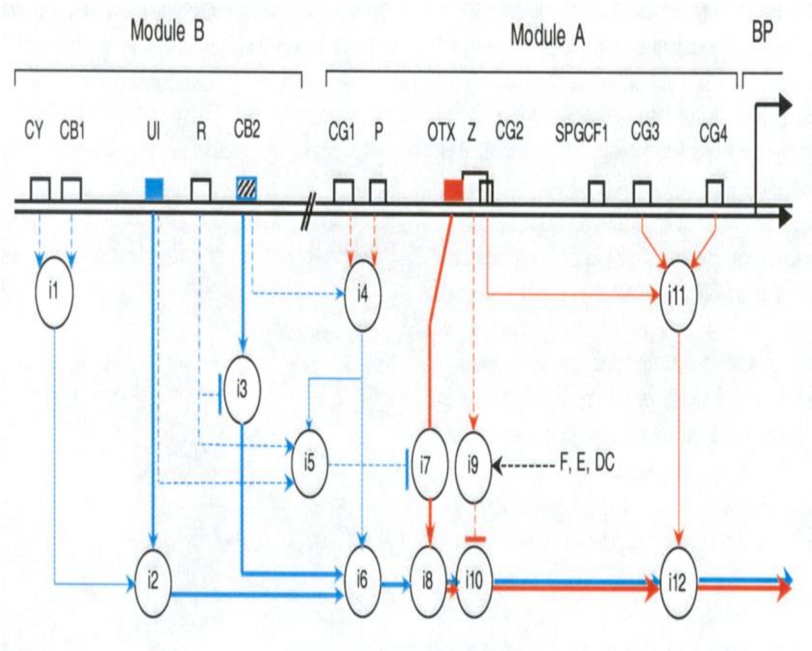
Drosophila

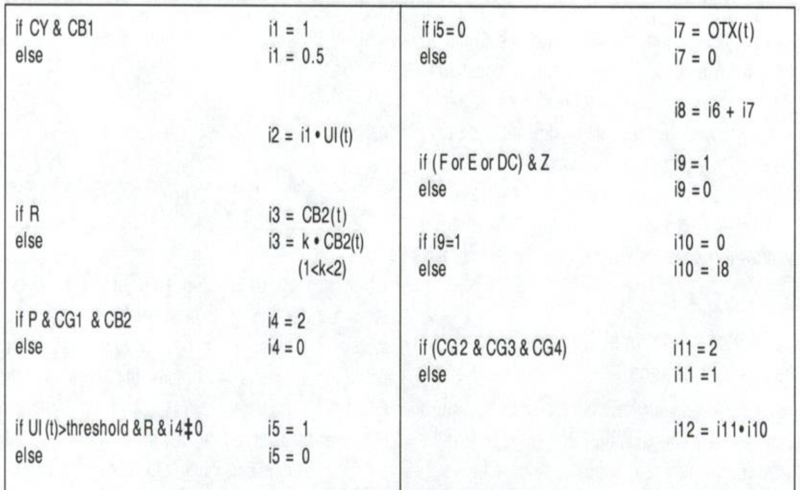




One gene, 30 years of study, 300 docs and postdocs

A Proposal for Nobel Prize

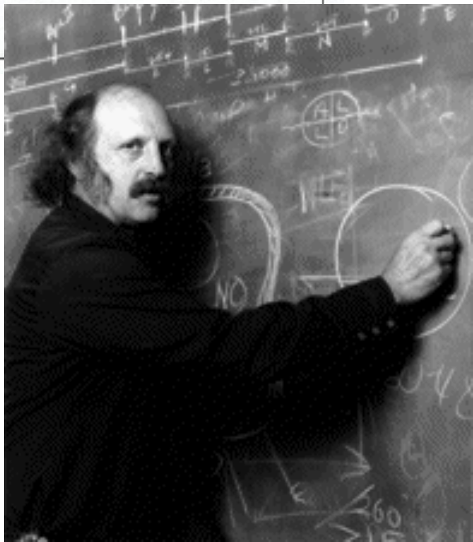




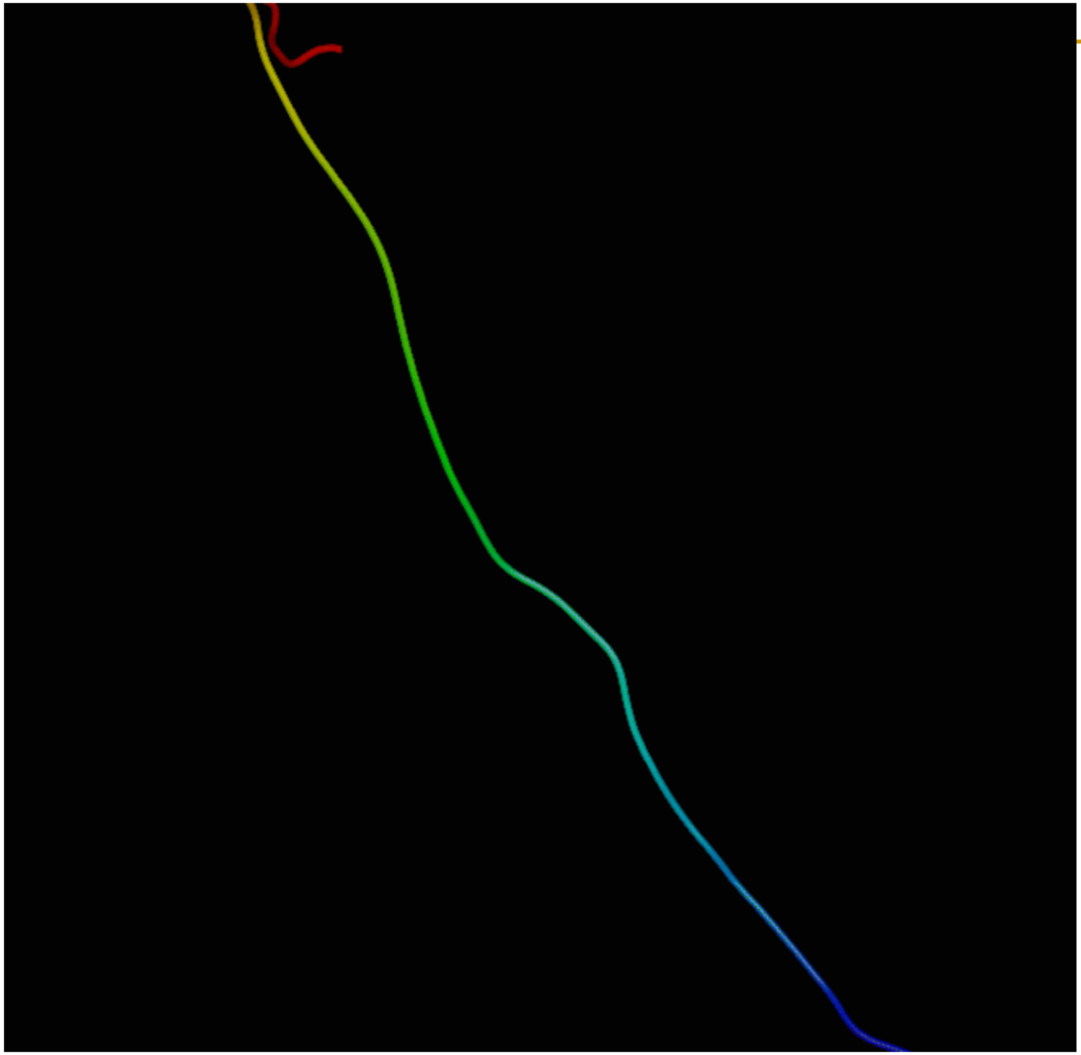
if CY & CB1	i1 = 1	if i5=0	i7 = OTX(t)
else	i1 = 0.5	else	i7 = 0
			i8 = i6 + i7
	i2 = i1 • UI(t)		
		if (F or E or DC) & Z	i9 = 1
		else	i9 = 0
if R	i3 = CB2(t)		
else	i3 = k • CB2(t)	if i9=1	i10 = 0
	(1<k<2)	else	i10 = i8
if P & CG1 & CB2	i4 = 2		
else	i4 = 0	if (CG2 & CG3 & CG4)	i11 = 2
		else	i11 =1
if UI (t)>threshold & R & i4‡0	i5 = 1		i12 = i11•i10
else	i5 = 0		
	i6 = i4 • (i2+i3)		

“Programs built into the DNA of every animal.”

Eric H. Davidson



Genomic Regulatory Systems



Prof. Istrail

The Protein Folding Problem

Statistical Mechanics
models

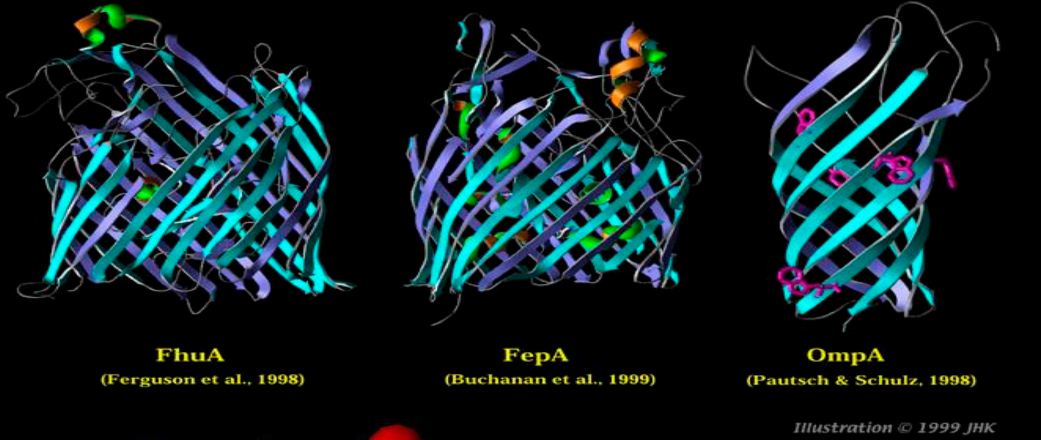
Mixed character of the problem :

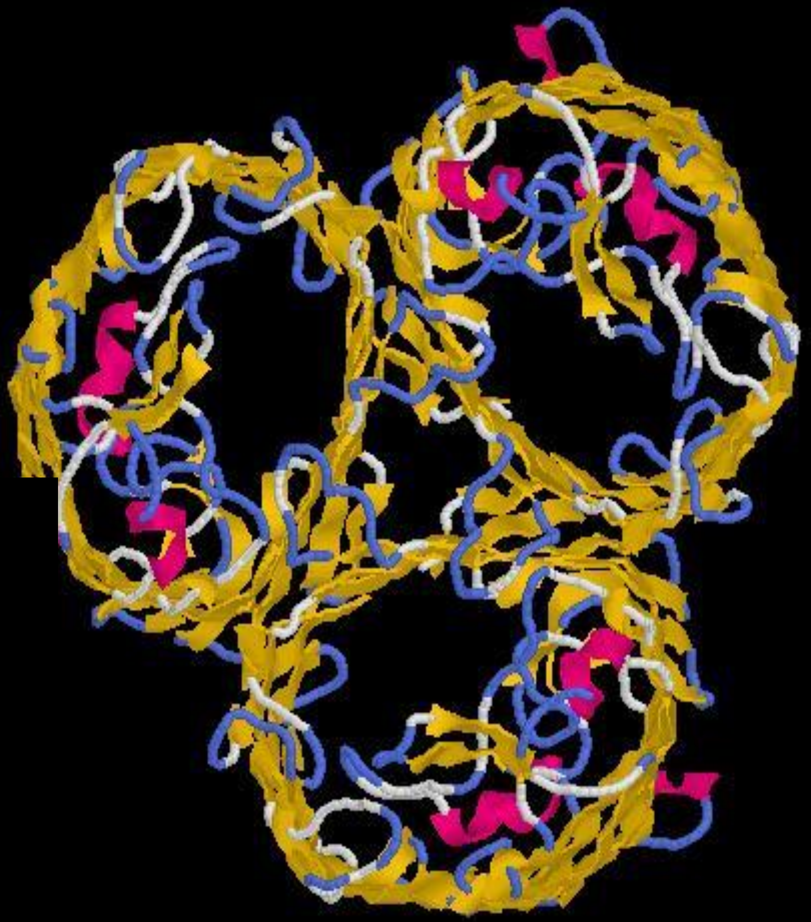
continuous mathematics -- geometry of surfaces &
discrete mathematics -- combinatorics of folds

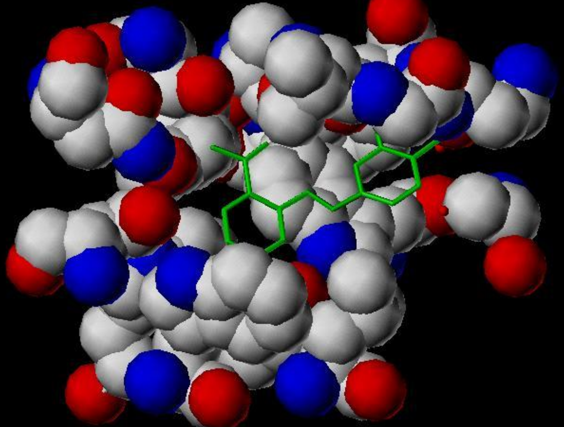
"The protein folding problem is three different problems: the folding code – the thermodynamic question of how a native structure results from the interatomic forces acting on an amino acid sequence; protein structure prediction – the computational problem of how to predict the native structure of a protein from its amino acid; and the folding speed (Levinthal's paradox) – the kinetic question of how a protein can fold so fast… Current knowledge of the folding code is sufficient to guide the successful design of new proteins and new materials. Current computer algorithms are now predicting the native structures of small simple proteins remarkable accurately, contributing to drug discovery and proteomics. Even once intractable Levinthal 'spuzzle now seems to have a very simple answer…"

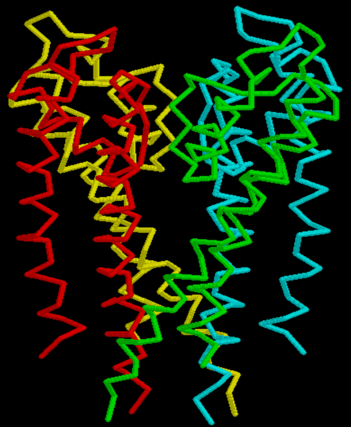
Ken Dill

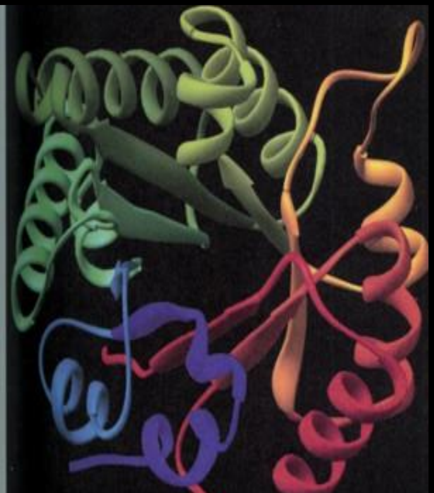
K. A. Dill, S. Banu Ozkan, T. R. Weikl, J. D. Chodera and V. A. Voelz. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, 17:2342-346, 2007.

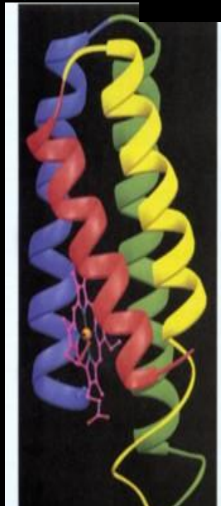

FhuA
(Ferguson et al., 1998)
FepA
(Buchanan et al., 1999)
OmpA
(Pautsch & Schulz, 1998)
Illustration © 1999 JHK

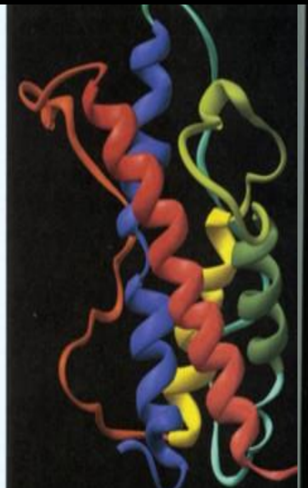




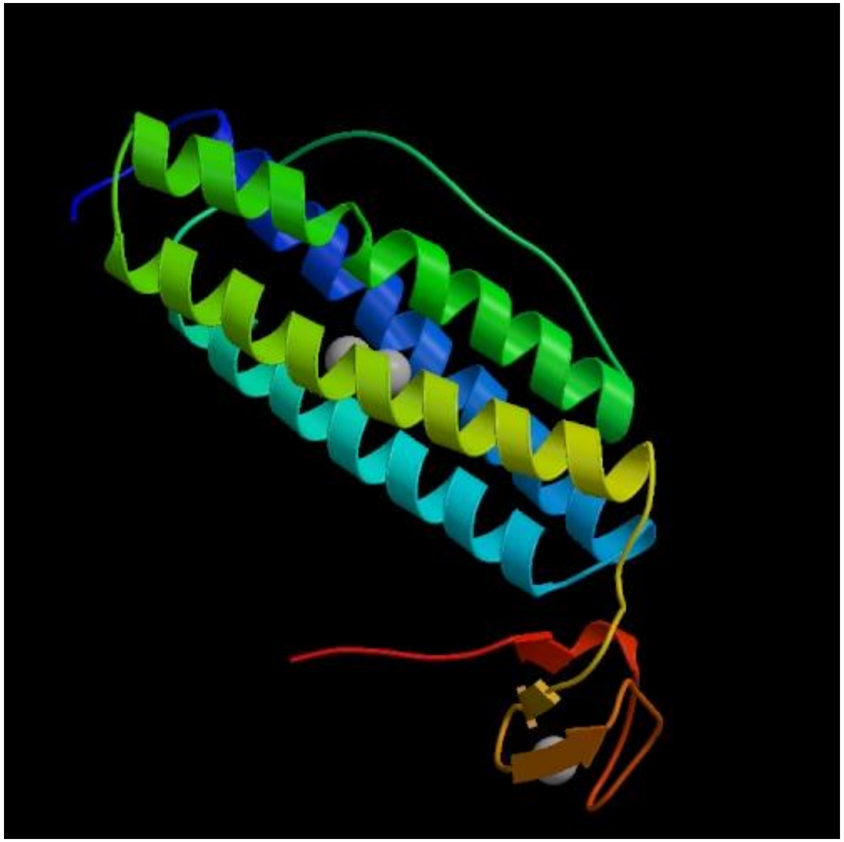




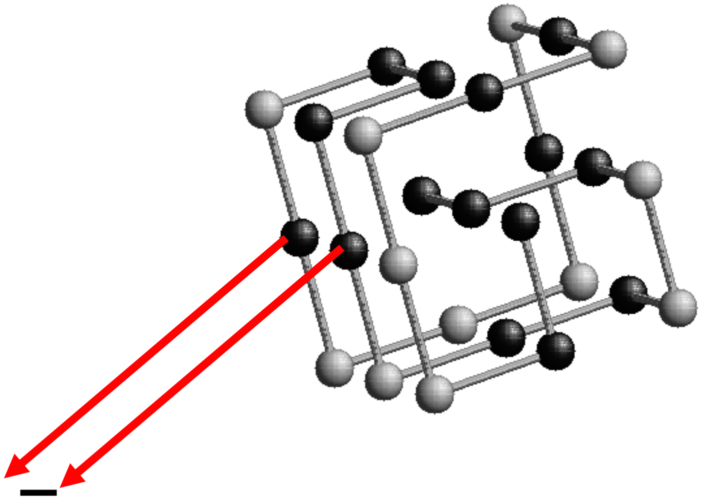


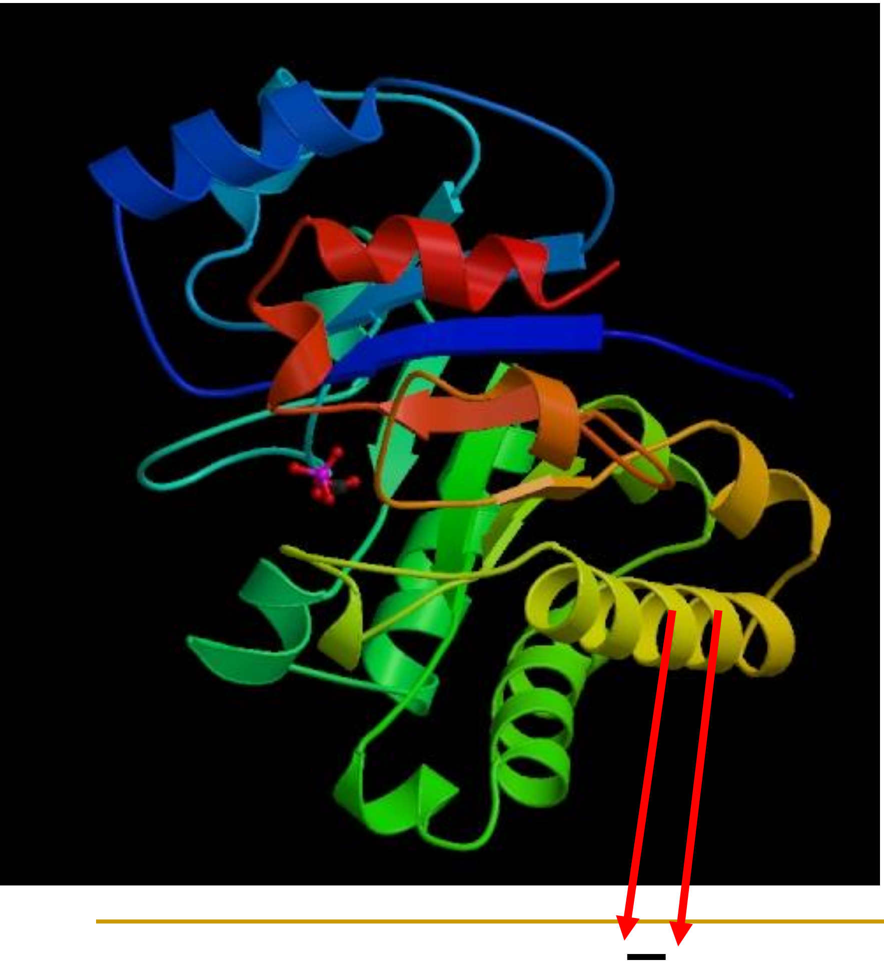


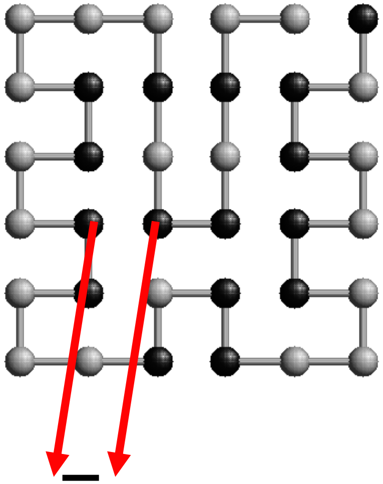
SELF-AVOIDING WALK



CONTACT

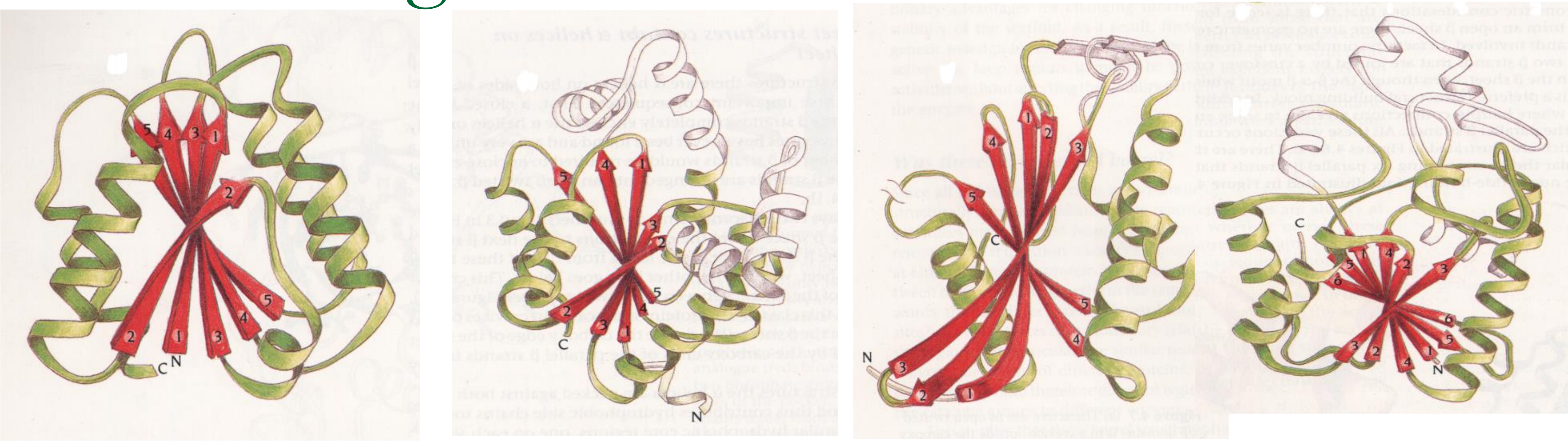






<= 4 Amstrongs

ASILOMAR: Protein Folding -- A Paradigm-Shift



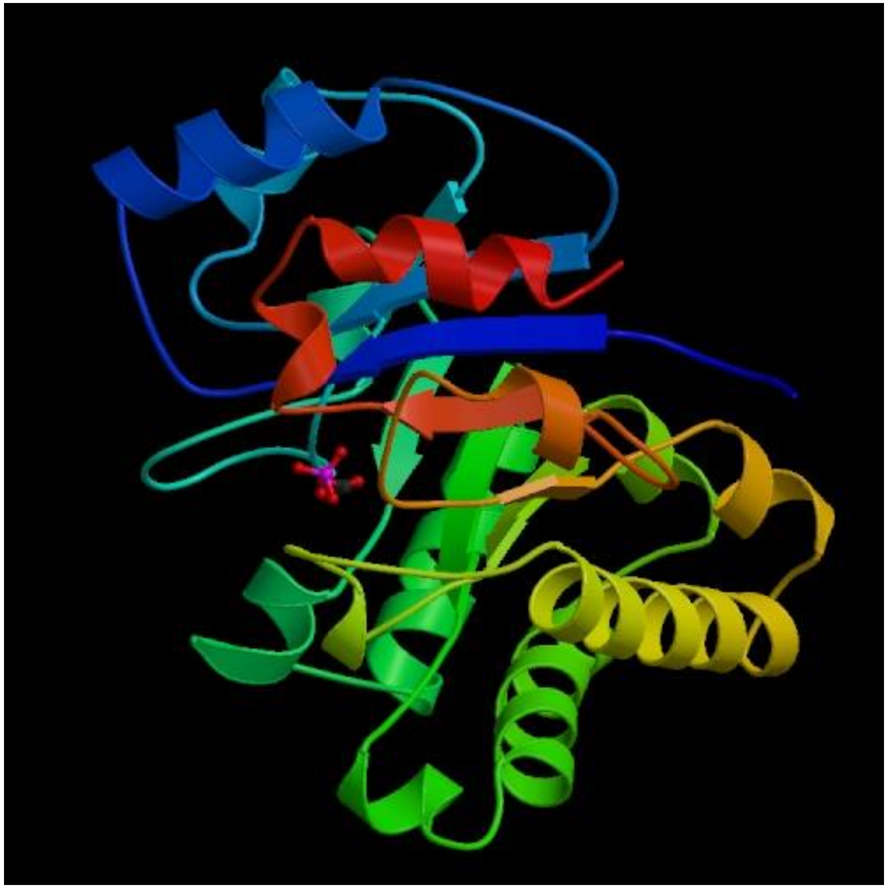
Structure Similarity

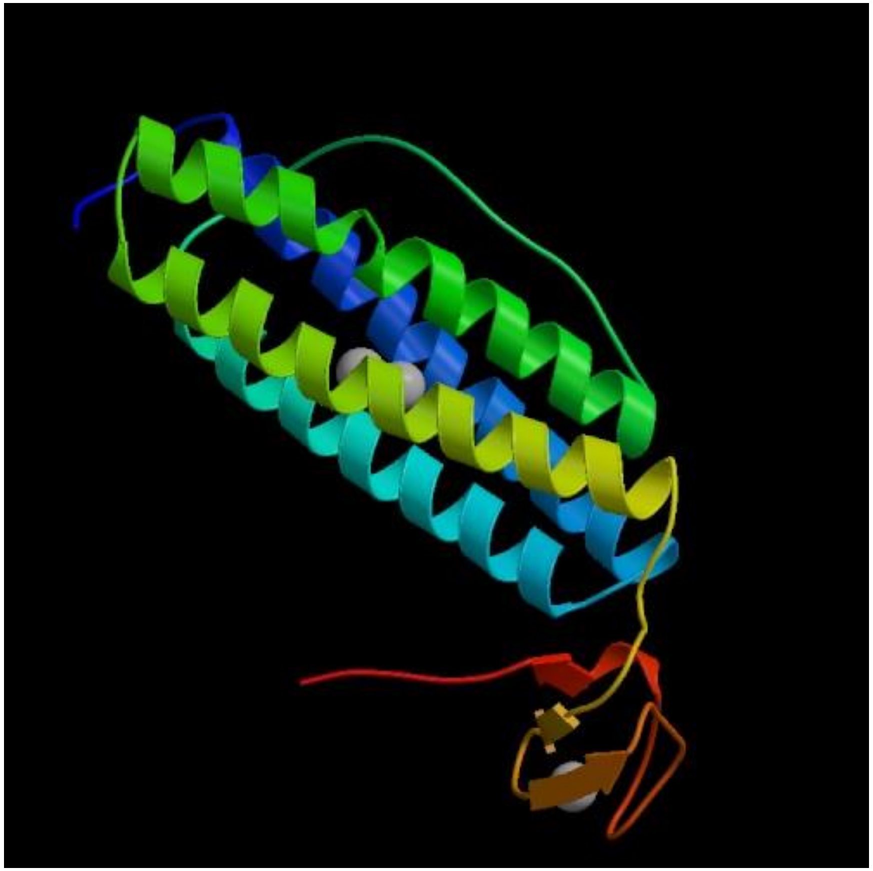
Structure Alignment

Fold Recognition

Fold Assignment

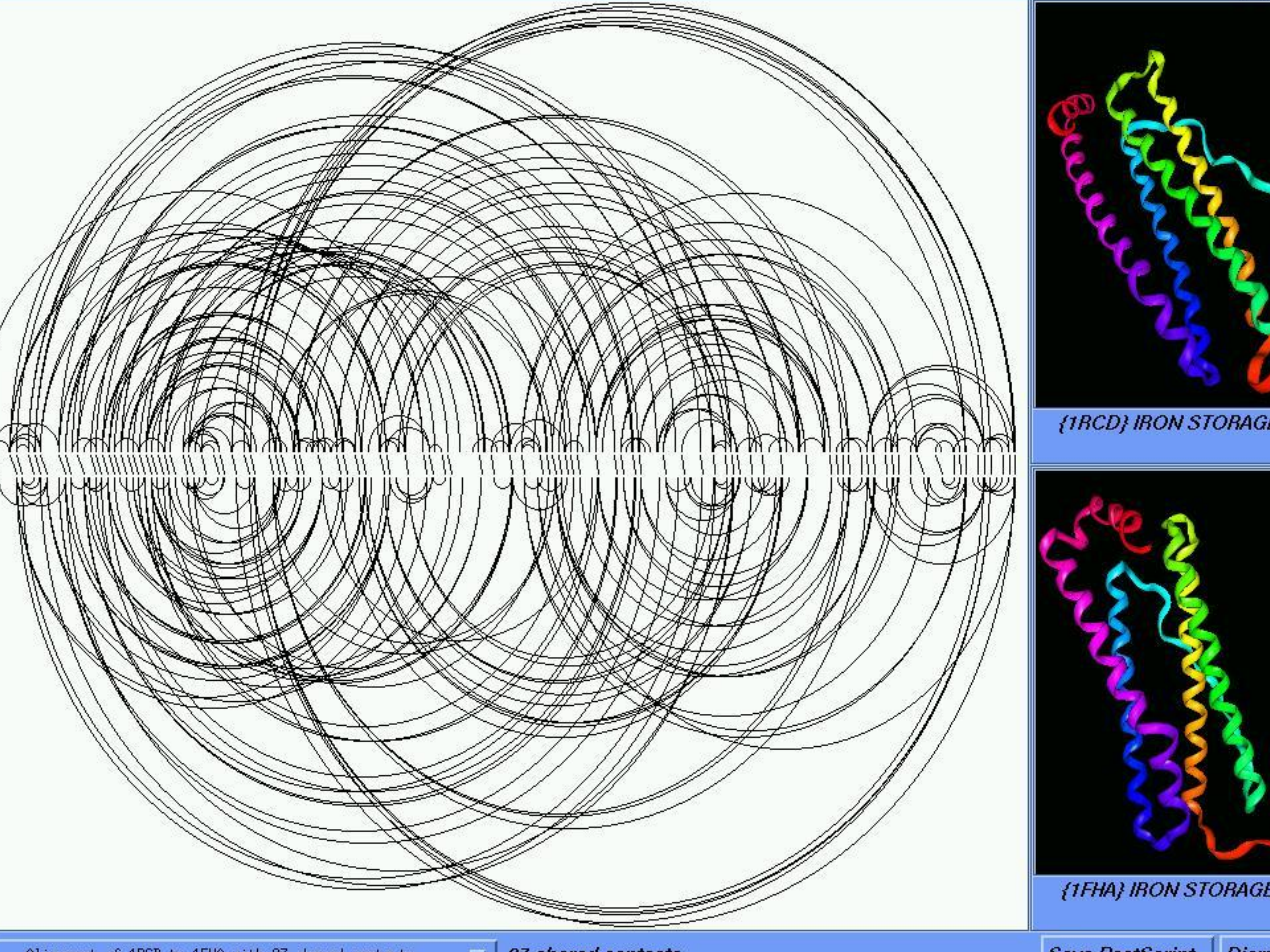
FOLD RECOGNITION





Measure of Protein Structure Similarity

- Root-mean-square distance (RMSD)
- Difference of the distance matrices (DDM)
- Contact Map Overlap (CMO)
- Various more or less ad hoc scoring schemes based on local secondary structure, hydrogen bonding pattern, burial status, interaction pattern



{1RCD} IRON STORAG
{1FHA} IRON STORAGE
Alignment of 1RCD to 1FHA with 87 shared contacts
87 shared contacts
Save PostScript
Disp

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin
- alpha-beta
- 8 structures
- up to 124 residues
- 15-30% sequence sin
- < 3Å RMSD

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin

- beta
- 8 structures
- up to 99 residues
- 35-90% sequence sin
- < 2Å RMSD

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin

- alpha-beta
- 11 structures
- up to 250 residues
- 30-90% sequence sin
- < 2Å RMSD

Skolnick Test

- Four Families
  1. Flavodoxin-like fold Che-Y related
  2. Plastocyanin
  3. TIM Barrel
  4. Ferratin

- alpha
- 6 structures
- up to 170 residues
- 7-70% sequence simi
- < 4Å RMSD