

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

```

Hsa_FAM46C consensus -hfvavlnwQDV-LldgvlsEviPIhGRGNPFLievqgpkdviqvVrsrLeeqgi-VhdvRL
61 80 90 100 110
Hsa_FAM46C NGSAAHGVVVDNLDGCGKLLILKVAPEATLGLVADVYCSNLSGSGKVKLLIK
Hsa_FAM46A NGSAAHGVVVDSDSLGYKGLGLILCADRGEGSLVTVKDVYDGLGSGKVKIKL
Hsa_FAM46B NGSAAHGVVHDSGSLGYKGLGLILVTVRESDGLTVKVAIVACIDPFGAGVRIITF
Hsa_FAM46D NGSAAHGVVASHNLSGSLGYKGLGLILVTVRESDGLTVKVAIVACIDPFGAGVRIITF
consensus nGSAAshvld-dGfvgKdIILf-vdi-gpaeafGLVkdVldGLdLdFvGvkeKf
121 130 140 150 160 170
Hsa_FAM46C VTLKEAVYGLNKKVCGDdSRSLISLSNKGKNNKLVKFDVSLRQKFVFVDFQIILDL
Hsa_FAM46A LTLKEAVYGVKMKVCGDdSRSLISLSNKGKNNKLVKFDVSLRQKFVFVDFQIILDL
Hsa_FAM46B LTLKEAVYGVKMKVCGDdSRSLISLSNKGKNNKLVKFDVSLRQKFVFVDFQIILDL
Hsa_FAM46D DIMEADVYGLNKKVCGDdSRSLISLSNKGKNNKLVKFDVSLRQKFVFVDFQIILDM
consensus ltlKEAVYGLNKKVCG-dsDRSLISLSN-gKKNvLKFVDVSLRQKFVFVDFQIILdL
181 190 200 210 220 230
Hsa_FAM46C VFYFDGCMNPFTEHFAFVTVIGCMVQDQAEAFDNGKLLAFNNFEIRGGGLKFSNNL
Hsa_FAM46A VFYFDGCMNPFTEHFAFVTVIGCMVQDQAEAFDNGKLLAFNNFEIRGGGLKFSNNL
Hsa_FAM46B LTLGQCSTFPMFAEFHFAFVTVIGCMVQDQAEAFDNGKLLAFNNFEIRGGGLKFSNNL
Hsa_FAM46D VYFSDGKNKALTEKSYVTVVAFLYDQDQAMTLGKLLAFNNFEIRGGGLKFSNNL
consensus LfYFDGcmnpnpe-fhFtvigfESmTGdGFqAfzHLq-rlaTFNFEIRGGGLKfYNNL
241 250 260 270 280 290
Hsa_FAM46C VADPAPDQDEIKETIEHNCMAFFIDPDILGQKLLIKNNFAEERER-IOYNNL
Hsa_FAM46A VADPAPAS-DEIKTQNNCFNFFIDPDILGQKLLIKNNFVGLDREK-IVYNNL
Hsa_FAM46B VADPAPDQDQDPAKAEHNCMAFFIDPDILGQKLLIKNNFAGGABAAR-ACVNNL
Hsa_FAM46D VADPAPCMSIRKIEHNCMAFFIDPDILGQKLLIKNNFAGGABAAR-ACVNNL
consensus VADPAPdQdEIKETIEHNCMAFFIDPDILGQKLLIKNNFAEERER-IOYNNL

```

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

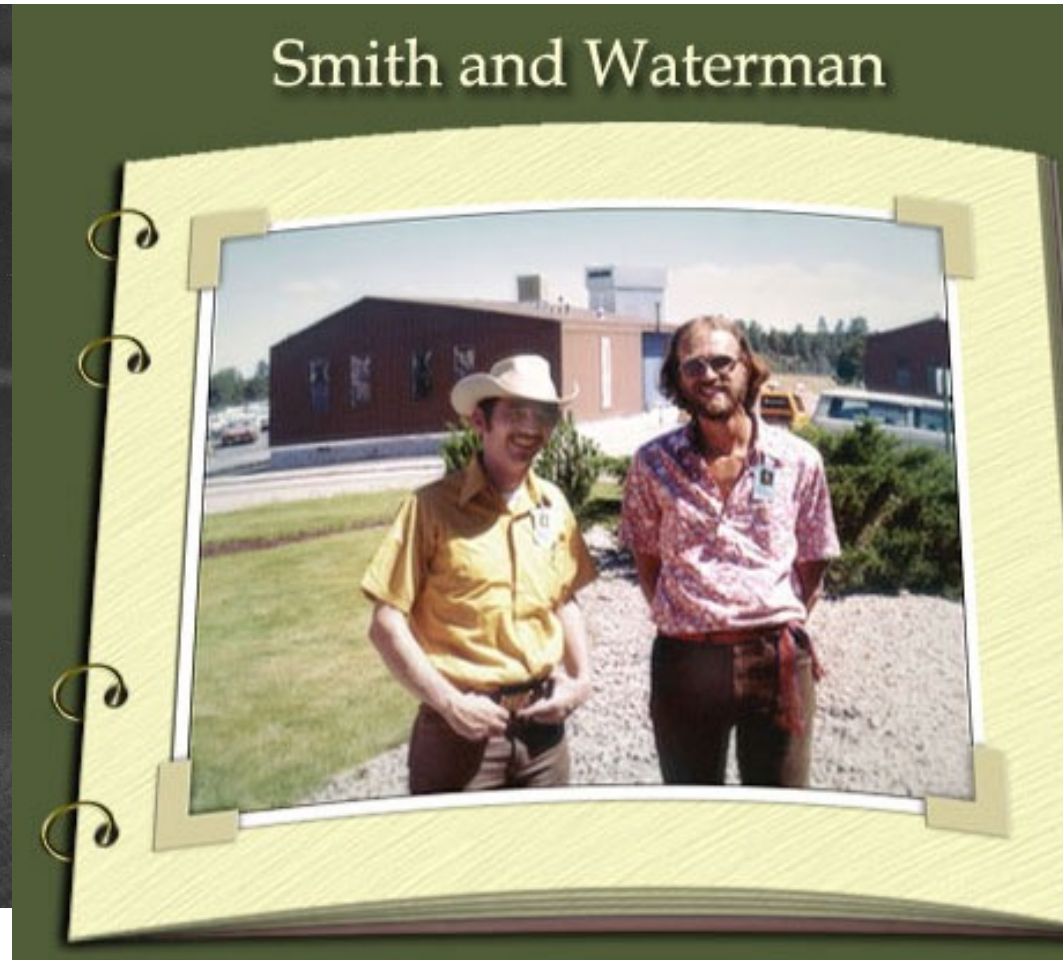
Problem 3. De-noising: how long an alignment needs to be non-random?

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		<ul style="list-style-type: none">• Linear discriminant analysis is a generalization of Fisher's linear discriminant^{[47][83]}• Fisher information, see also scoring algorithm also known as the Fisher scoring, and Minimum Fisher information, a variational principle which, when applied with the proper constraints needed to match 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 to the central limit theorem 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 the "fundamental paradox of inverse probability" as the source of the "confusion between statistical terms which refer to the true value of a parameter estimated, with the actual value arrived at by estimation, with the result being subject to error."^[86]• Fisher's permutation test• Fisher's inequality^[87]• Sufficient statistic, when a statistic is <i>sufficient</i> with respect to a statistical model and its associated unknown parameter if there is no statistic that can be calculated from the same sample providing additional information as to the value of the parameter"^[88]• Fisher's noncentral hypergeometric distribution, a generalization of the hypergeometric distribution, where sampling probabilities are modified by weight factors.• Student's t-distribution, widely used in statistics.^{[89][90]}• The concept of an ancillary statistic and the notion (the ancillary principle) that one should condition on ancillary statistics.
	17 February 1890		
	London, England		
Died	29 July 1962 (aged 72)		
	Adelaide , South Australia , Australia		



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 scaffolds?
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?

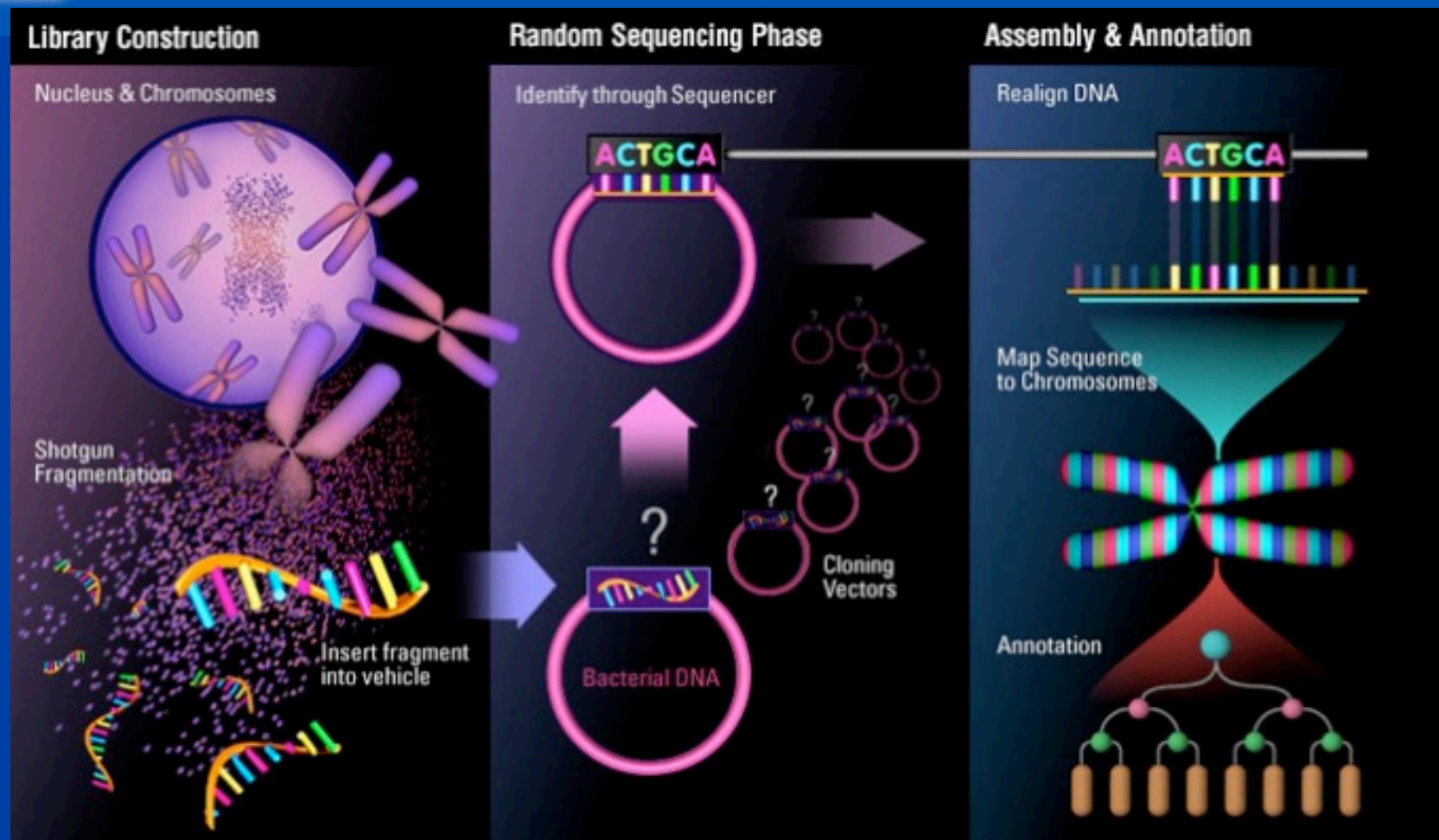
Problems we need to solve along the way

Problem 1. Genome Assembly Algorithms

Problem 2. Poisson statistics for DNA and Genome Assembly

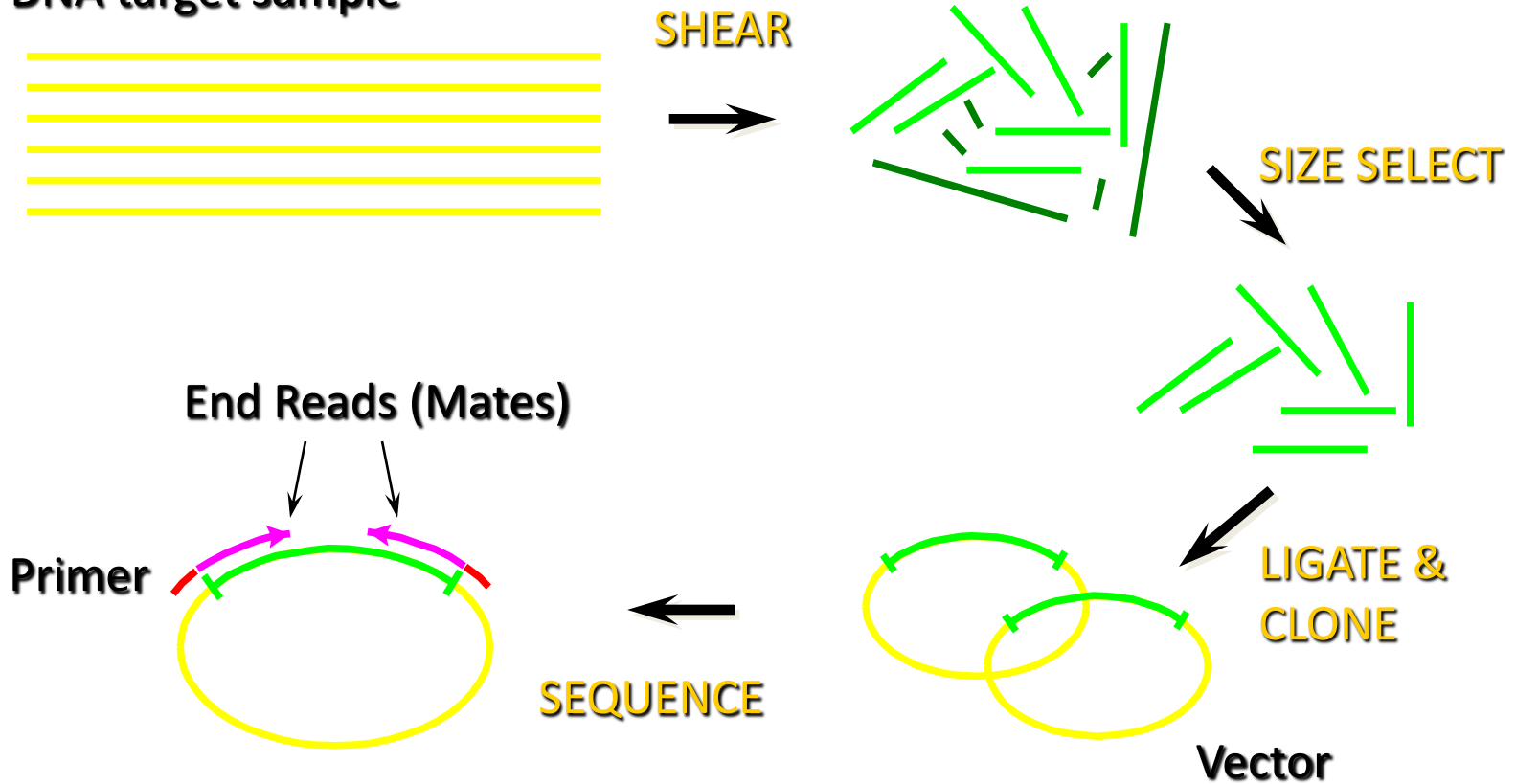
Problem 3. Ham Smith's DNA Lab with no windows

Whole Genome Shotgun Sequencing

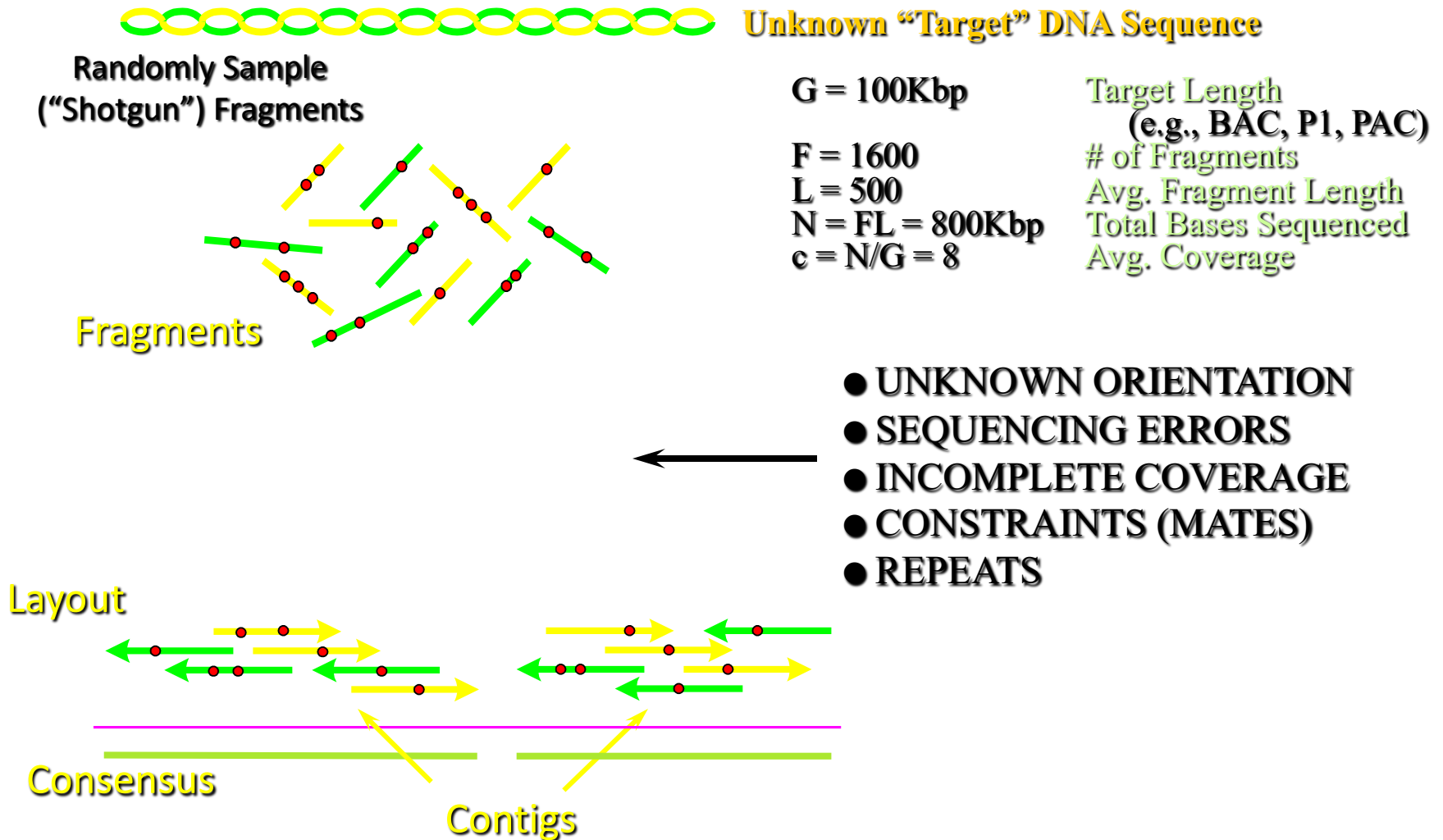


Shotgun DNA Sequencing (Technology)

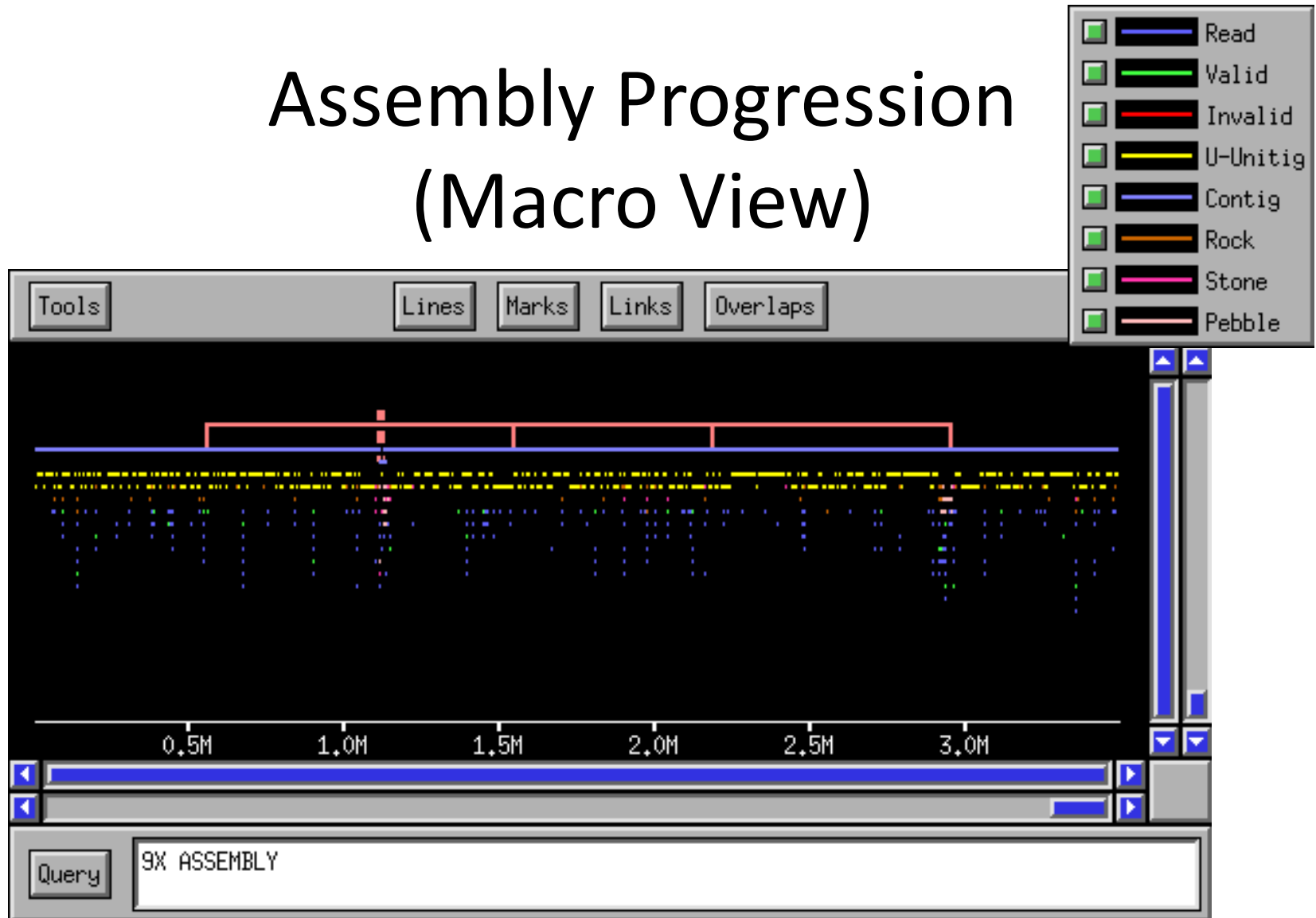
DNA target sample



Shotgun DNA Sequencing (Computation)



Assembly Progression (Macro View)

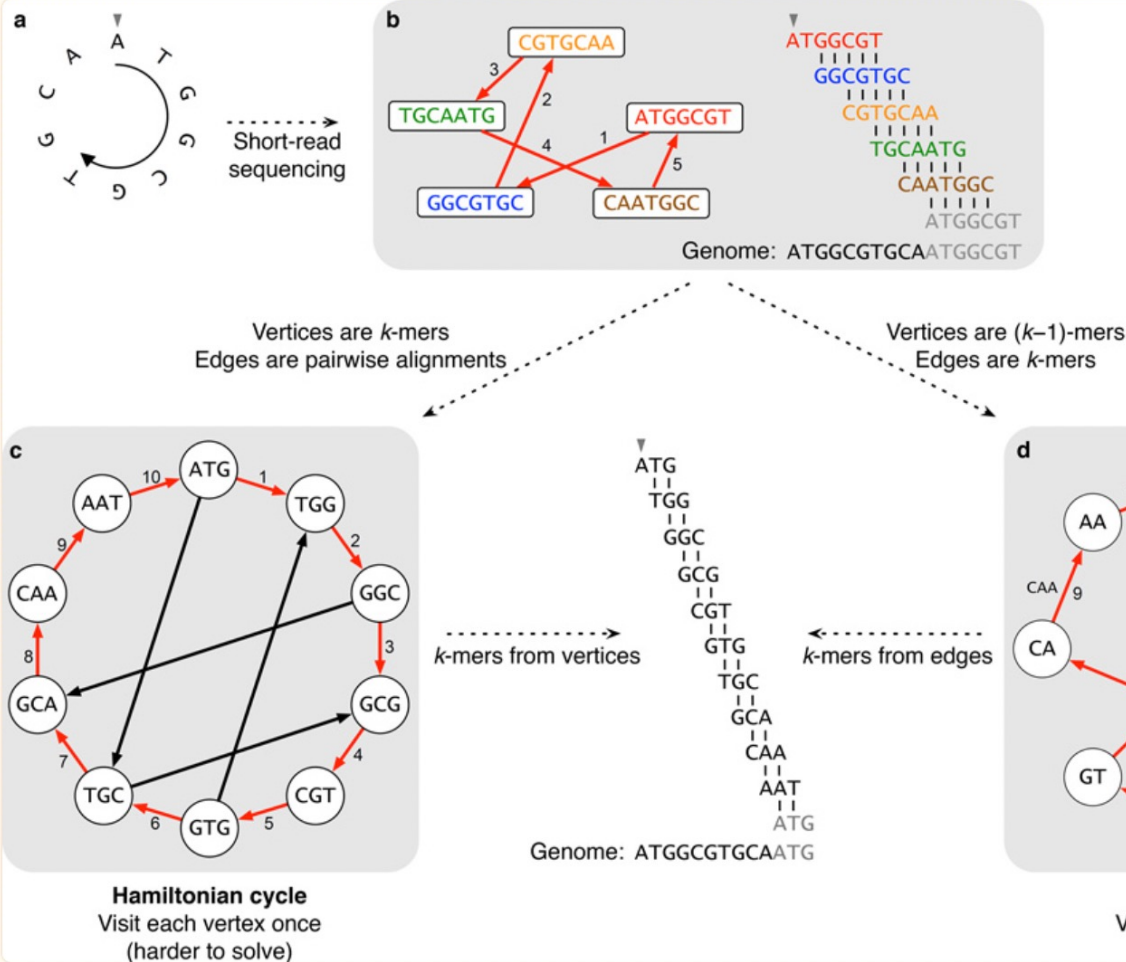


Siméon Denis Poisson

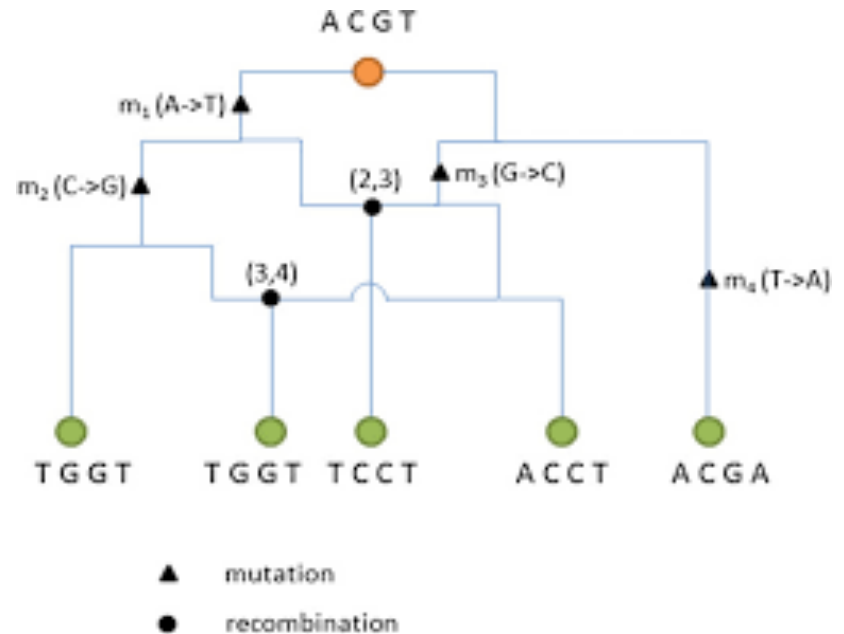
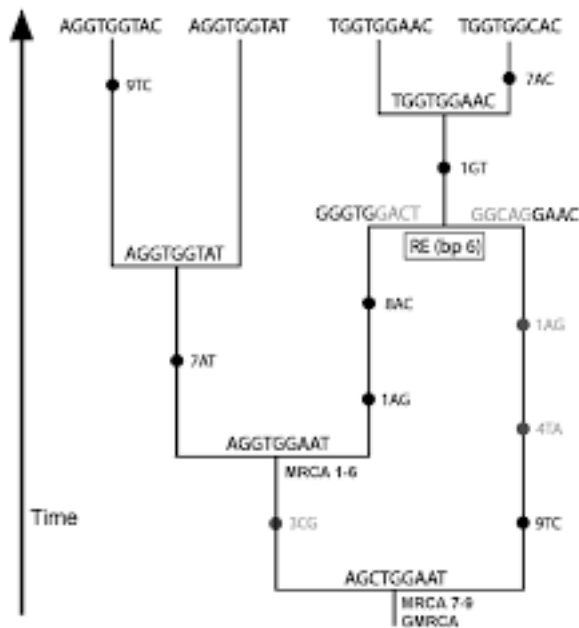


Born	21 June 1781 Pithiviers, Kingdom of France (present-day Loiret)
Died	25 April 1840 (aged 58) Sceaux, Hauts-de-Seine, Kingdom of France
Alma mater	École Polytechnique
Known for	Poisson process Poisson equation Poisson kernel Poisson distribution Poisson limit theorem Poisson bracket Poisson algebra 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



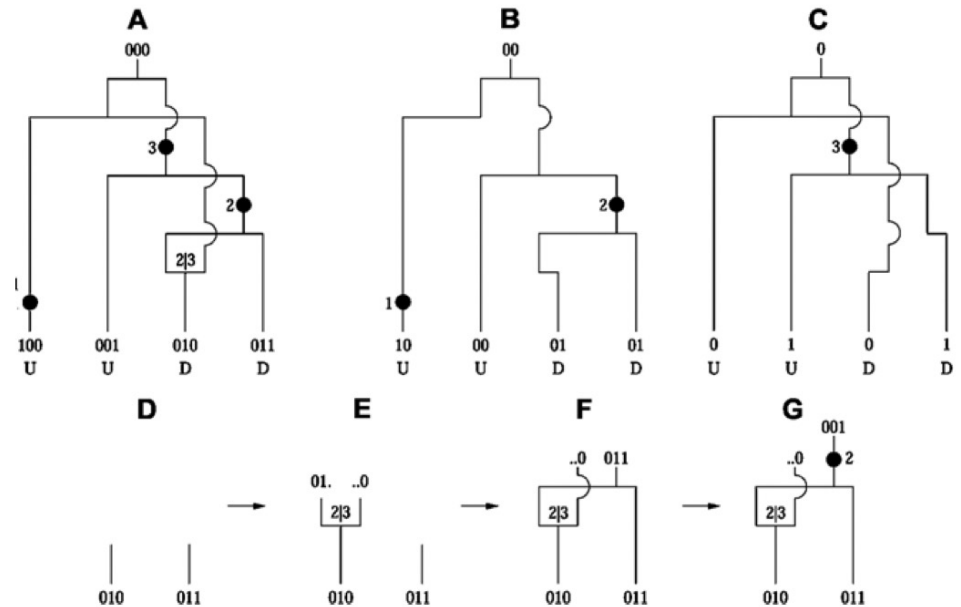
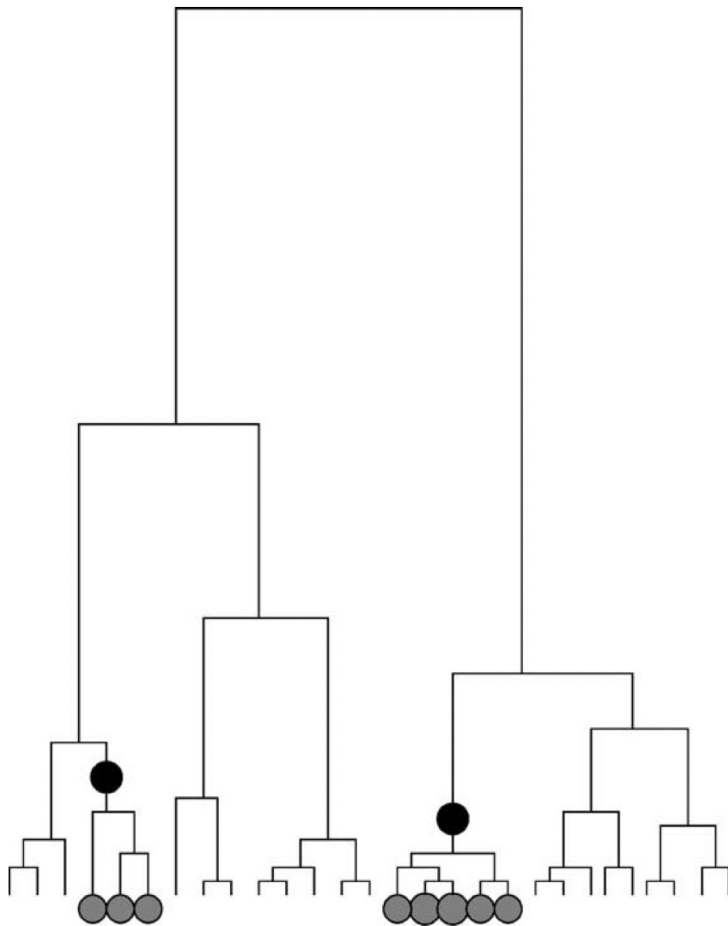
Ch. 3 Recombination and Ancestral Recombination Graphs (ARG) Algorithms



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

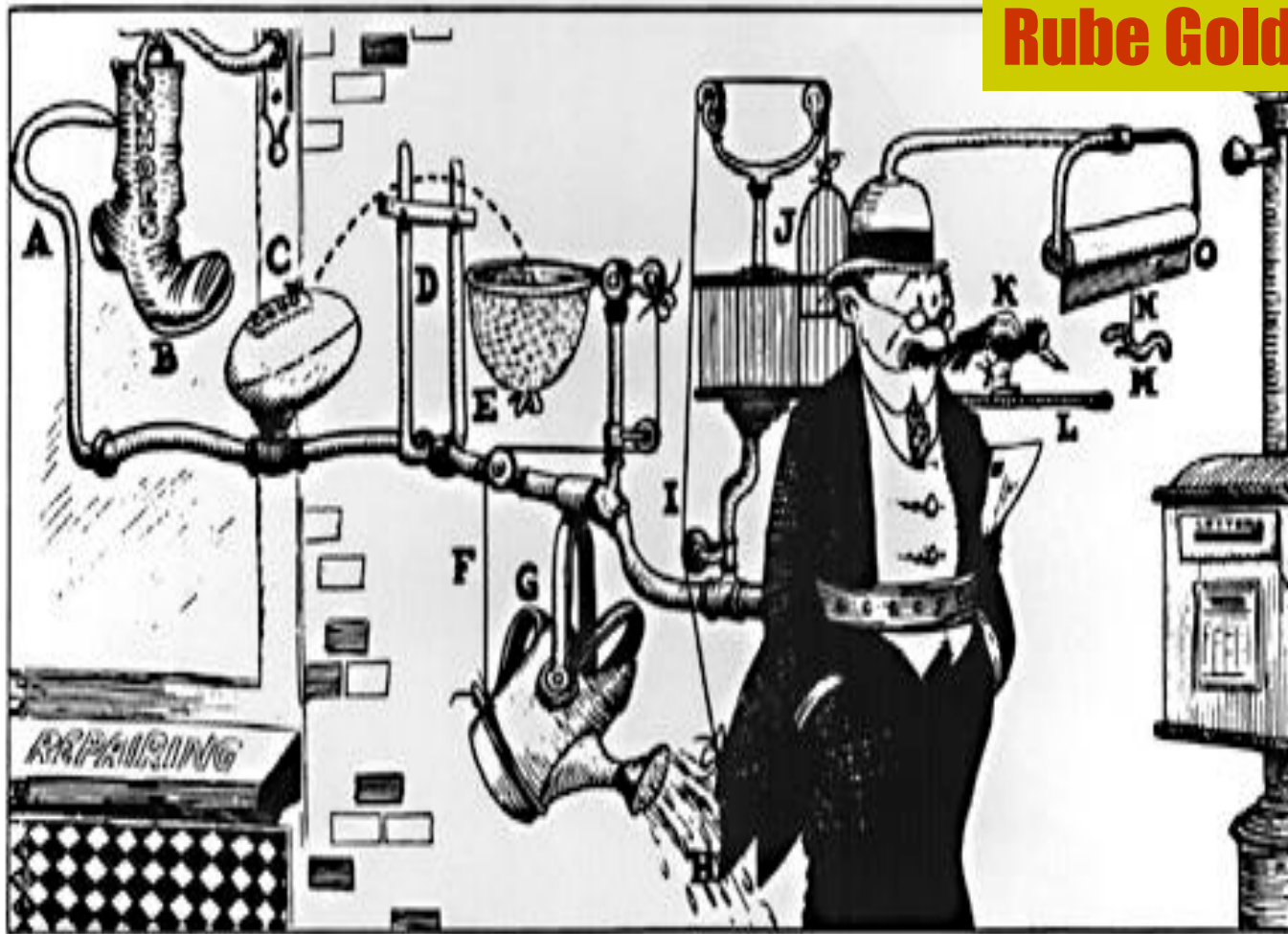


Born	23 March 1749 Beaumont-en-Auge , Normandy, Kingdom of France
Died	5 March 1827 (aged 77) Paris , Kingdom of France
Alma mater	University of Caen
Known for	show
Scientific career	
Fields	Astronomy and Mathematics

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**

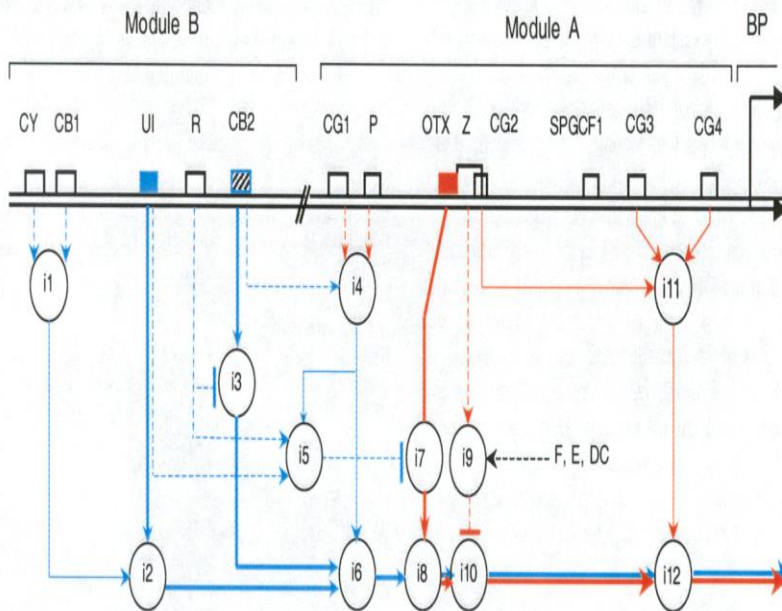
A composite image consisting of two parts. On the left is a photograph of a dark, spiny sea urchin resting on a light-colored, textured surface. On the right is a detailed, colored illustration of a housefly, shown from a side profile. The fly has a brown and black segmented body, large red eyes, and transparent wings with visible veins.

The diagram illustrates a complex signaling pathway network. Key components and interactions include:

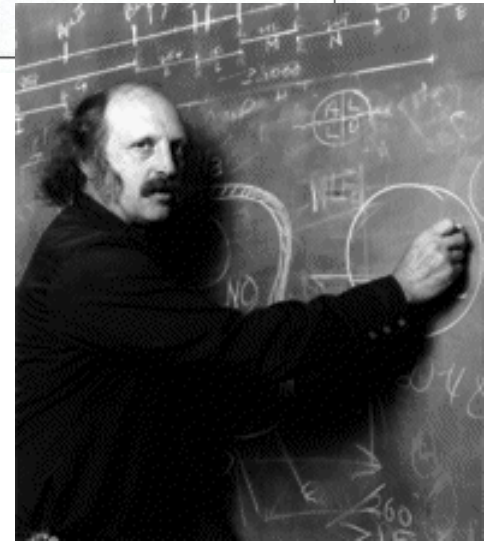
- Left Panel (Pink background):** Shows the EGF signaling pathway. EGF binds to EGF-R, activating a cascade involving Ras, Raf, MEK, and ERK. Other components like TGF-α, p187, p190, p195, p200, p205, p210, p215, p220, p225, p230, p235, p240, p245, p250, p255, p260, p265, p270, p275, p280, p285, p290, p295, p300, p305, p310, p315, p320, p325, p330, p335, p340, p345, p350, p355, p360, p365, p370, p375, p380, p385, p390, p395, p400, p405, p410, p415, p420, p425, p430, p435, p440, p445, p450, p455, p460, p465, p470, p475, p480, p485, p490, p495, p500, p505, p510, p515, p520, p525, p530, p535, p540, p545, p550, p555, p560, p565, p570, p575, p580, p585, p590, p595, p600, p605, p610, p615, p620, p625, p630, p635, p640, p645, p650, p655, p660, p665, p670, p675, p680, p685, p690, p695, p700, p705, p710, p715, p720, p725, p730, p735, p740, p745, p750, p755, p760, p765, p770, p775, p780, p785, p790, p795, p800, p805, p810, p815, p820, p825, p830, p835, p840, p845, p850, p855, p860, p865, p870, p875, p880, p885, p890, p895, p900, p905, p910, p915, p920, p925, p930, p935, p940, p945, p950, p955, p960, p965, p970, p975, p980, p985, p990, p995, p1000, p1005, p1010, p1015, p1020, p1025, p1030, p1035, p1040, p1045, p1050, p1055, p1060, p1065, p1070, p1075, p1080, p1085, p1090, p1095, p1100, p1105, p1110, p1115, p1120, p1125, p1130, p1135, p1140, p1145, p1150, p1155, p1160, p1165, p1170, p1175, p1180, p1185, p1190, p1195, p1200, p1205, p1210, p1215, p1220, p1225, p1230, p1235, p1240, p1245, p1250, p1255, p1260, p1265, p1270, p1275, p1280, p1285, p1290, p1295, p1300, p1305, p1310, p1315, p1320, p1325, p1330, p1335, p1340, p1345, p1350, p1355, p1360, p1365, p1370, p1375, p1380, p1385, p1390, p1395, p1400, p1405, p1410, p1415, p1420, p1425, p1430, p1435, p1440, p1445, p1450, p1455, p1460, p1465, p1470, p1475, p1480, p1485, p1490, p1495, p1500, p1505, p1510, p1515, p1520, p1525, p1530, p1535, p1540, p1545, p1550, p1555, p1560, p1565, p1570, p1575, p1580, p1585, p1590, p1595, p1600, p1605, p1610, p1615, p1620, p1625, p1630, p1635, p1640, p1645, p1650, p1655, p1660, p1665, p1670, p1675, p1680, p1685, p1690, p1695, p1700, p1705, p1710, p1715, p1720, p1725, p1730, p1735, p1740, p1745, p1750, p1755, p1760, p1765, p1770, p1775, p1780, p1785, p1790, p1795, p1800, p1805, p1810, p1815, p1820, p1825, p1830, p1835, p1840, p1845, p1850, p1855, p1860, p1865, p1870, p1875, p1880, p1885, p1890, p1895, p1900, p1905, p1910, p1915, p1920, p1925, p1930, p1935, p1940, p1945, p1950, p1955, p1960, p1965, p1970, p1975, p1980, p1985, p1990, p1995, p2000, p2005, p2010, p2015, p2020, p2025, p2030, p2035, p2040, p2045, p2050, p2055, p2060, p2065, p2070, p2075, p2080, p2085, p2090, p2095, p2100, p2105, p2110, p2115, p2120, p2125, p2130, p2135, p2140, p2145, p2150, p2155, p2160, p2165, p2170, p2175, p2180, p2185, p2190, p2195, p2200, p2205, p2210, p2215, p2220, p2225, p2230, p2235, p2240, p2245, p2250, p2255, p2260, p2265, p2270, p2275, p2280, p2285, p2290, p2295, p2300, p2305, p2310, p2315, p2320, p2325, p2330, p2335, p2340, p2345, p2350, p2355, p2360, p2365, p2370, p2375, p2380, p2385, p2390, p2395, p2400, p2405, p2410, p2415, p2420, p2425, p2430, p2435, p2440, p2445, p2450, p2455, p2460, p2465, p2470, p2475, p2480, p2485, p2490, p2495, p2500, p2505, p2510, p2515, p2520, p2525, p2530, p2535, p2540, p2545, p2550, p2555, p2560, p2565, p2570, p2575, p2580, p2585, p2590, p2595, p2600, p2605, p2610, p2615, p2620, p2625, p2630, p2635, p2640, p2645, p2650, p2655, p2660, p2665, p2670, p2675, p2680, p2685, p2690, p2695, p2700, p2705, p2710, p2715, p2720, p2725, p2730, p2735, p2740, p2745, p2750, p2755, p2760, p2765, p2770, p2775, p2780, p2785, p2790, p2795, p2800, p2805, p2810, p2815, p2820, p2825, p2830, p2835, p2840, p2845, p2850, p2855, p2860, p2865, p2870, p2875, p2880, p2885, p2890, p2895, p2900, p2905, p2910, p2915, p2920, p2925, p2930, p2935, p2940, p2945, p2950, p2955, p2960, p2965, p2970, p2975, p2980, p2985, p2990, p2995, p3000, p3005, p3010, p3015, p3020, p3025, p3030, p3035, p3040, p3045, p3050, p3055, p3060, p3065, p3070, p3075, p3080, p3085, p3090, p3095, p3100, p3105, p3110, p3115, p3120, p3125, p3130, p3135, p3140, p3145, p3150, p3155, p3160, p3165, p3170, p3175, p3180, p3185, p3190, p3195, p3200, p3205, p3210, p3215, p3220, p3225, p3230, p3235, p3240, p3245, p3250, p3255, p3260, p3265, p3270, p3275, p3280, p3285, p3290, p3295, p3300, p3305, p3310, p3315, p3320, p3325, p3330, p3335, p3340, p3345, p3350, p3355, p3360, p3365, p3370, p3375, p3380, p3385, p3390, p3395, p3400, p3405, p3410, p3415, p3420, p3425, p3430, p3435, p3440, p3445, p3450, p3455, p3460, p3465, p3470, p3475, p3480, p3485, p3490, p3495, p3500, p3505, p3510, p3515, p3520, p3525, p3530, p3535, p3540, p3545, p3550, p35

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

A Proposal for Nobel Prize



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



“Programs built into the DNA of every animal.”

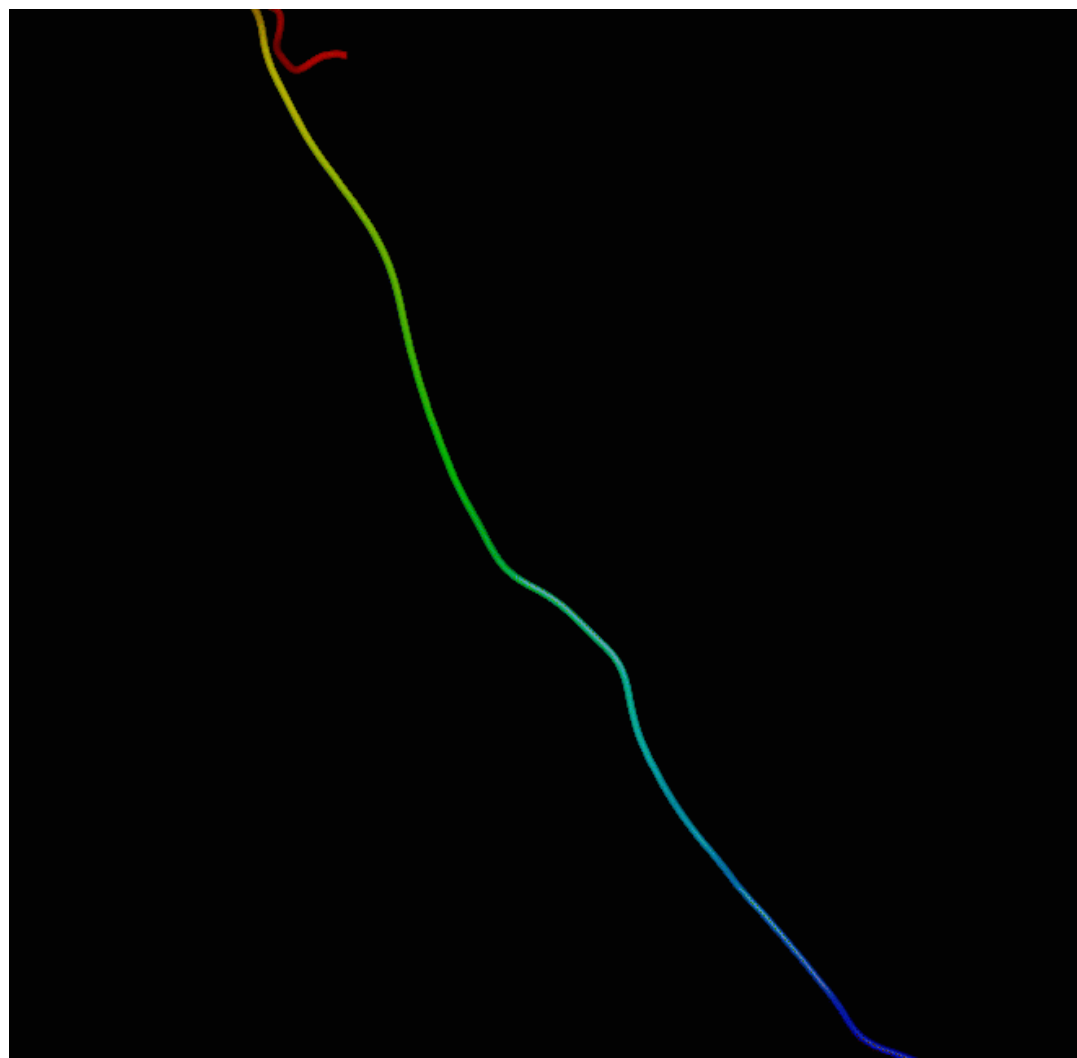
Eric H. Davidson

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

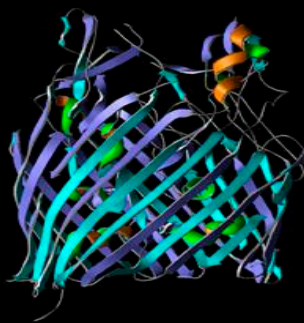
“The protein folding problem is three different problems: the folding code – the thermodynamic question of how a native structure results from the interatomic forces acting on an amino acid sequence; protein structure prediction – the computational problem of how to predict the native structure of a 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’s puzzle 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)

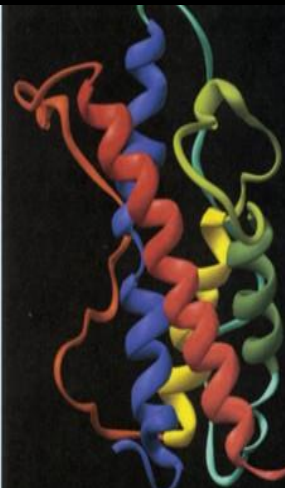
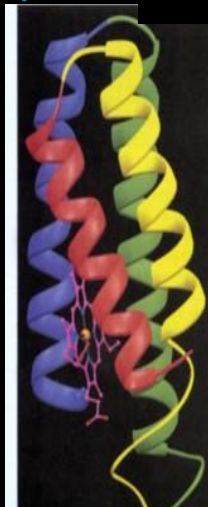
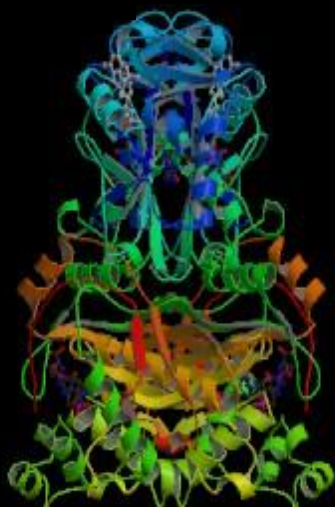
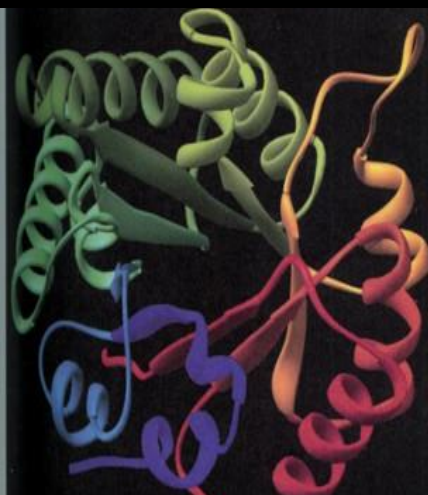
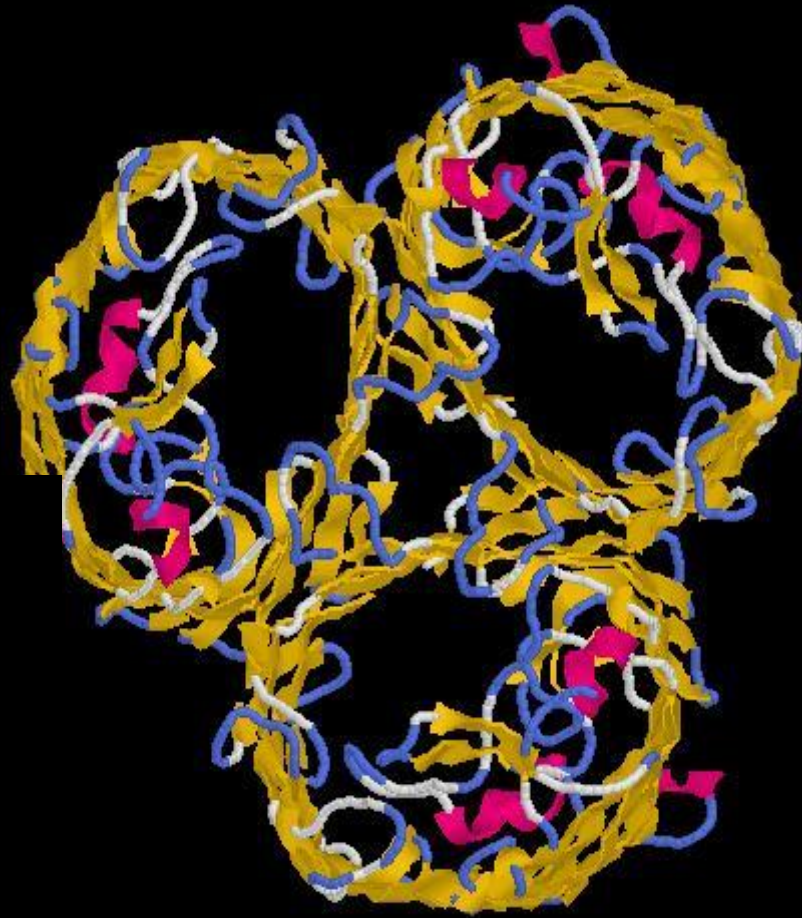
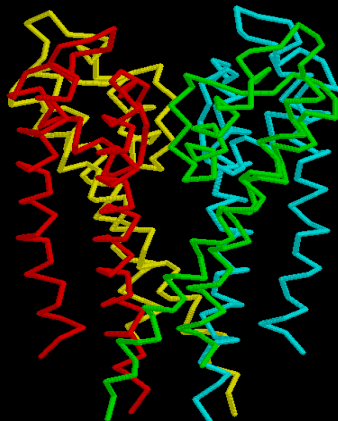
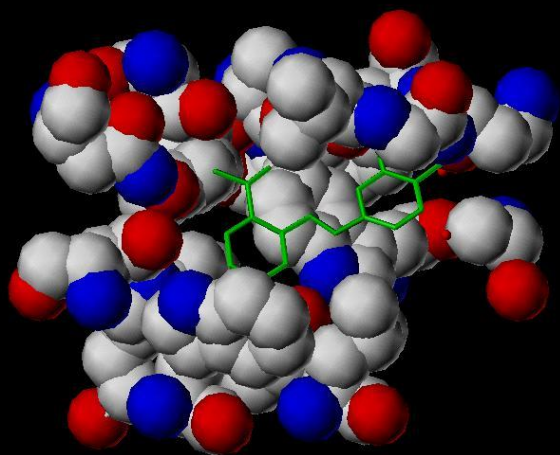


FepA
(Buchanan et al., 1999)

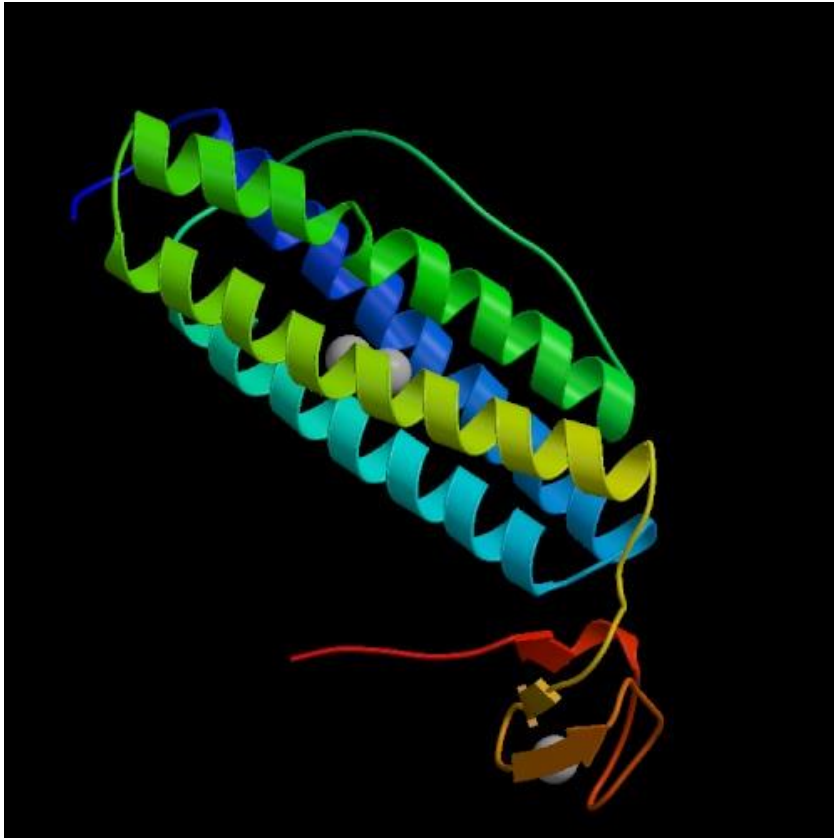


OmpA
(Pautsch & Schulz, 1998)

Illustration © 1999 JHK

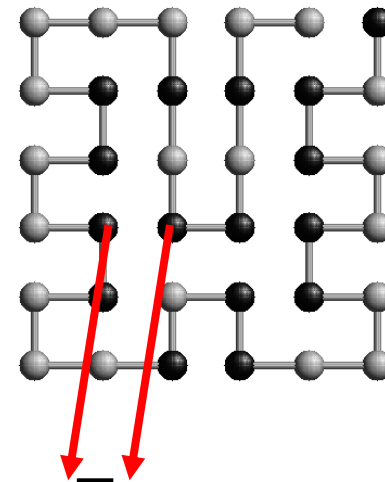
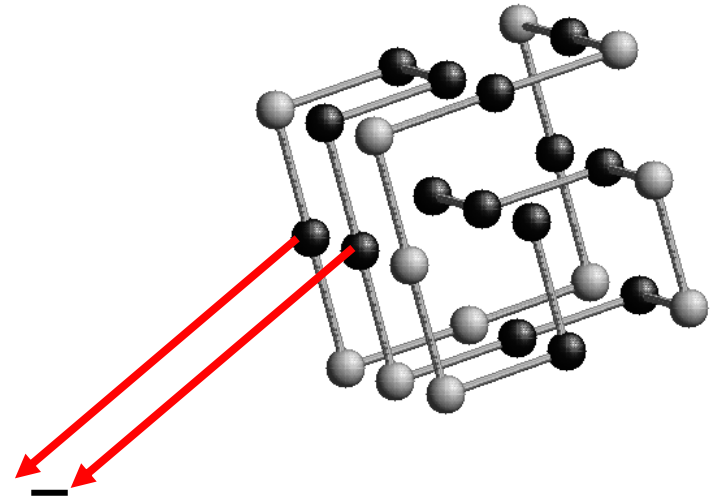
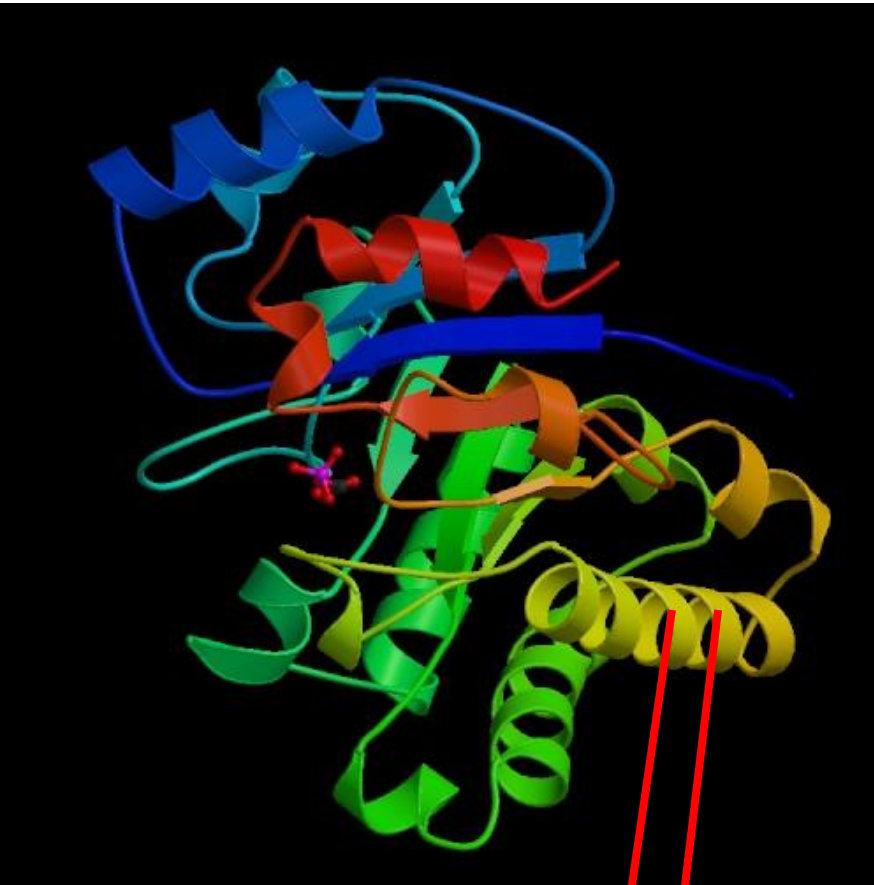


SELF-AVOIDING WALK



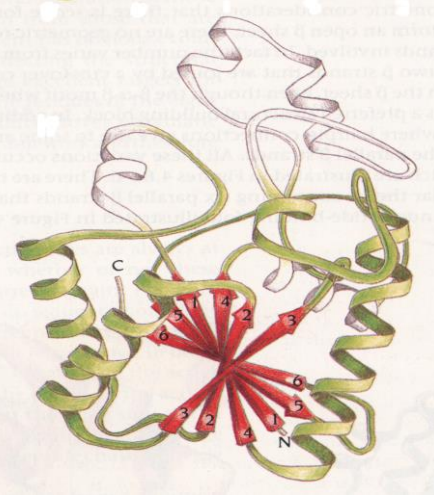
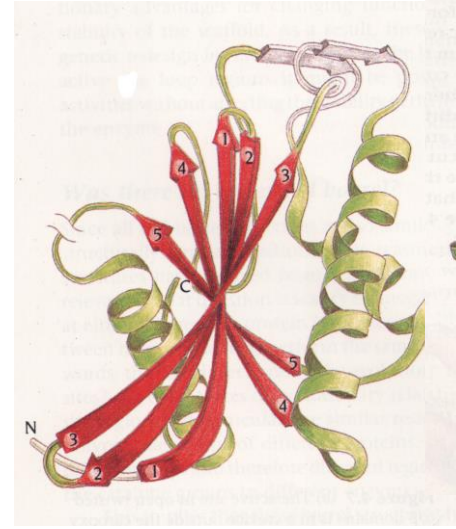
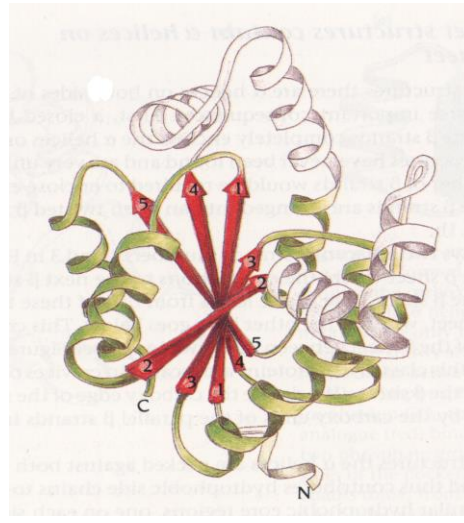
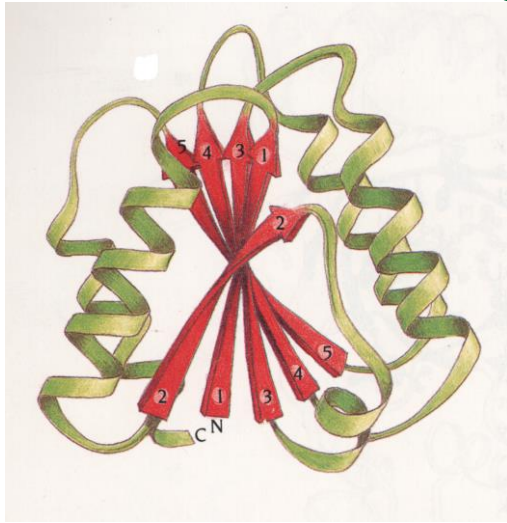


CONTACT



≤ 4 Amstrongs

ASILOMAR: Protein Folding -- A Paradigm-Shift



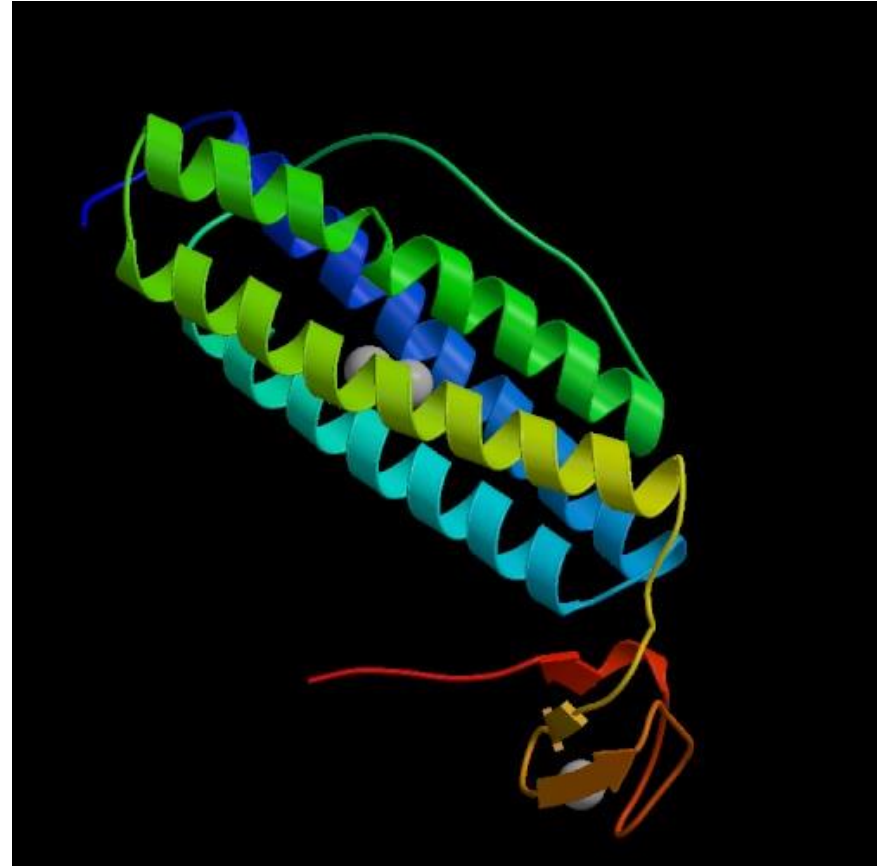
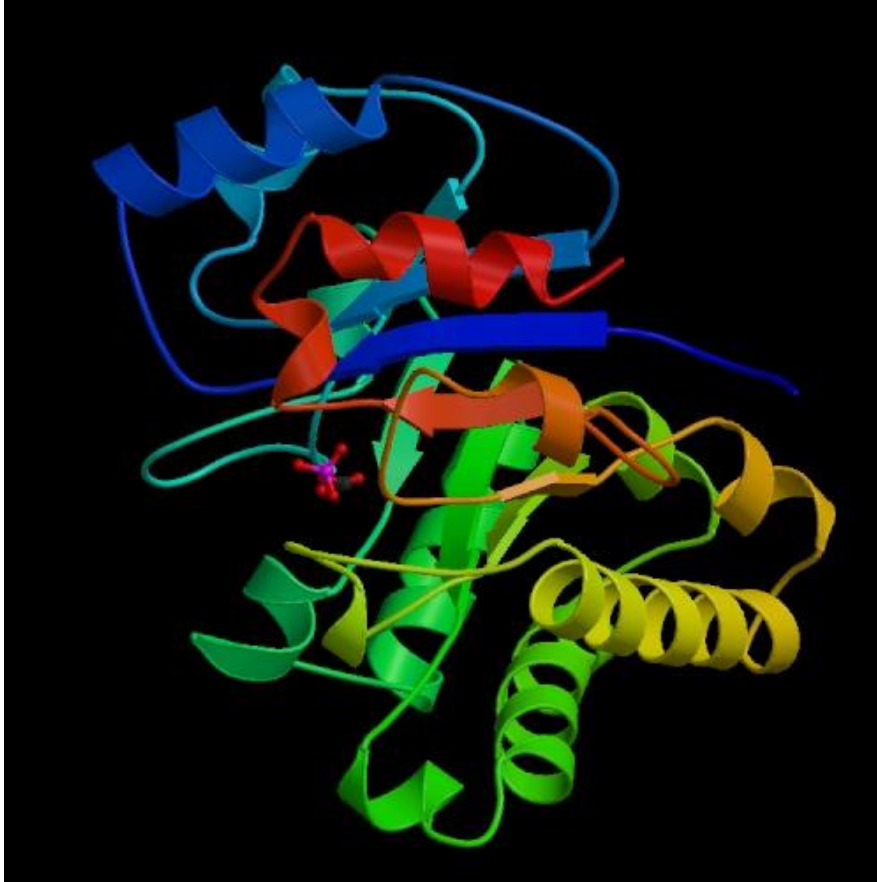
Structure Similarity

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
-

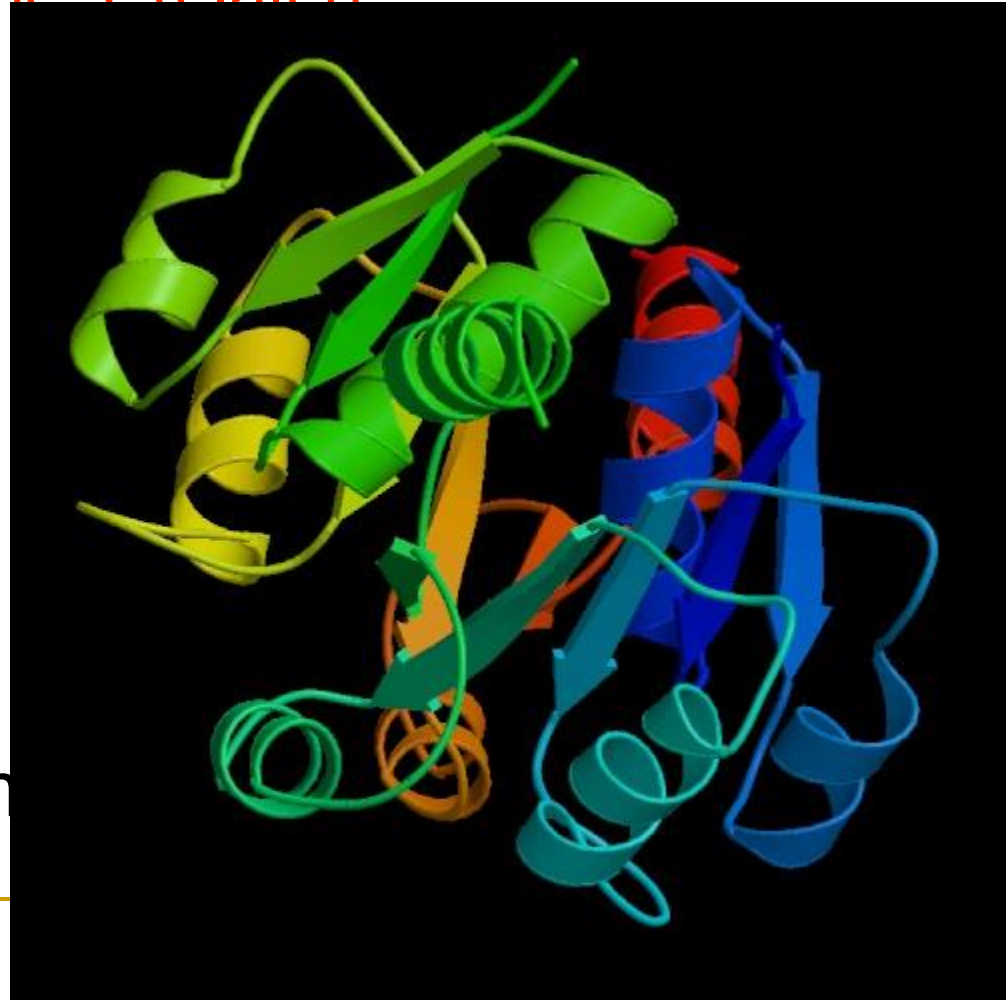


Skolnick Test

■ Four Families

- 1 Flavodoxin-like fold Che-Y related
- 2 Plastocyanin
- 3 TIM Barrel
- 4 Ferritin

- alpha-beta
- 8 structures
- up to 124 residues
- 15-30% sequence similarity
- $< 3\text{\AA}$ RMSD



Skolnick Test

■ Four Families

- 1 Flavodoxin-like fold Che-Y related
- 2 **Plastocyanin**
- 3 TIM Barrel
- 4 Ferritin

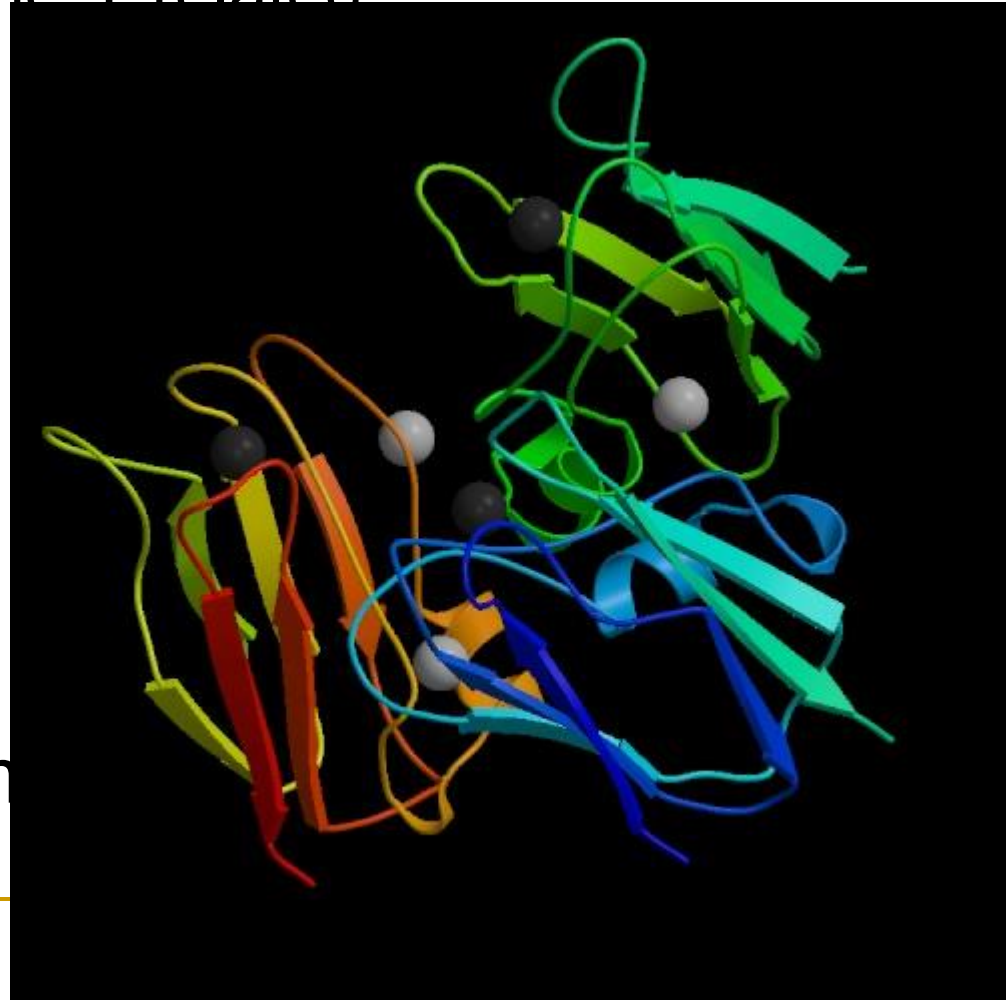
■ beta

■ 8 structures

■ up to 99 residues

■ 35-90% sequence sim

■ $< 2\text{\AA}$ RMSD



Skolnick Test

■ Four Families

- 1 Flavodoxin-like fold Che-Y related
- 2 Plastocyanin
- 3 **TIM Barrel**
- 4 Ferritin

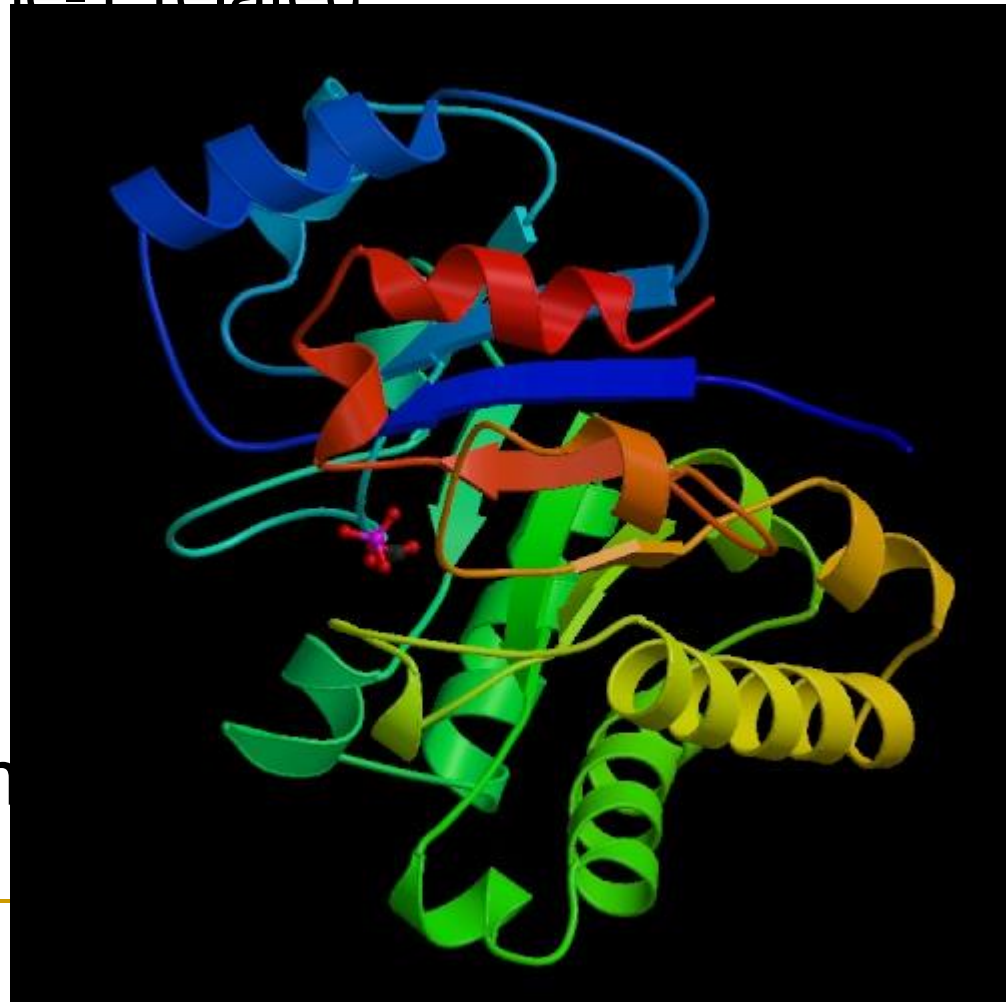
■ alpha-beta

■ 11 structures

■ up to 250 residues

■ 30-90% sequence sim

■ $< 2\text{\AA}$ RMSD



Skolnick Test

■ Four Families

- 1 Flavodoxin-like fold Che-Y related
- 2 Plastocyanin
- 3 TIM Barrel
- 4 **Ferritin**

- alpha
- 6 structures
- up to 170 residues
- 7-70% sequence similarity
- $< 4\text{\AA}$ RMSD

