Studying Macromolecular Motions in a Database Framework:

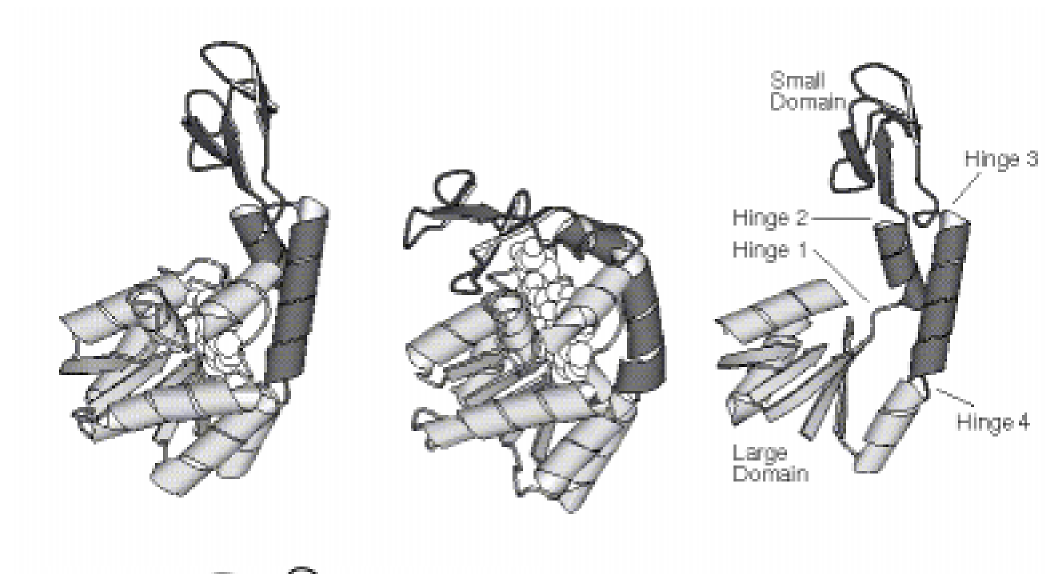
From Structure To Sequence

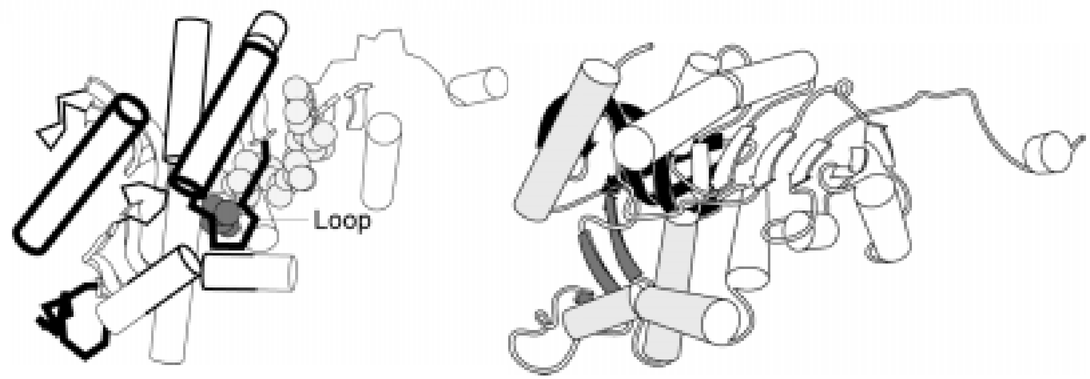
Mark Gerstein

Yale U.

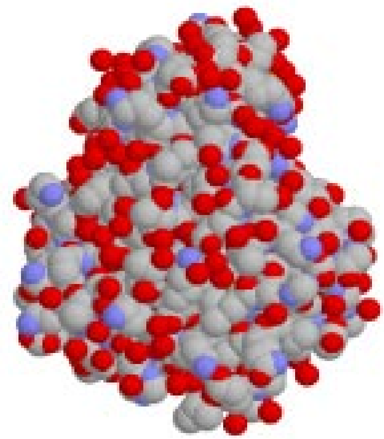
Macromolecular Motions

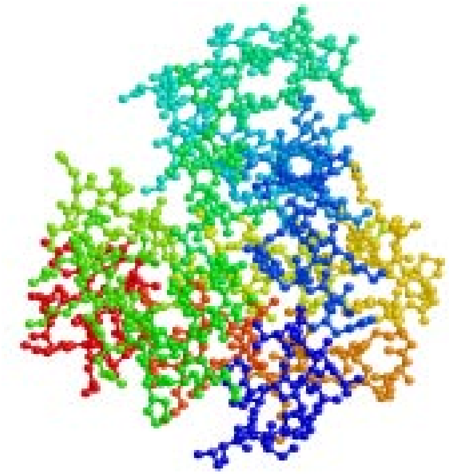
- What are they?
  - ◊ Proteins, Nucleic Acids (Hammerhead)
  - ◊ Sidechains (trivial), Loops (LDH), Domains (ADK), Subunits (Hb)
  - ◊ When a Ligand Binds: Open, Closed
- Essential link between structure and function
  - ◊ catalysis, regulation, transport, formation of assemblies, and cellular locomotion
- A complicated biological phenomena that can be studied in quantitative detail
  - ◊ changes in thousands of atomic coordinates

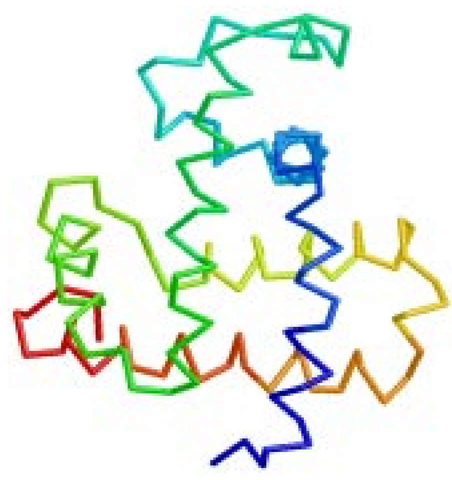


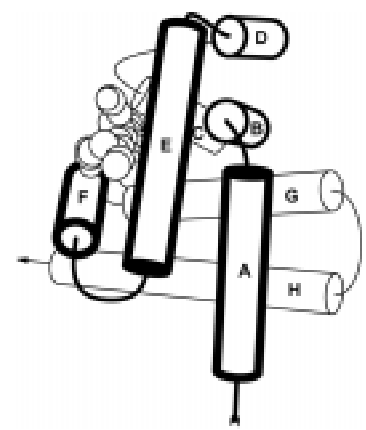


Depicting Protein Structure: Sperm Whale Myoglobin









Studying Macromolecular Motions in a Database Framework: from Structure to Sequence

1Motions Database

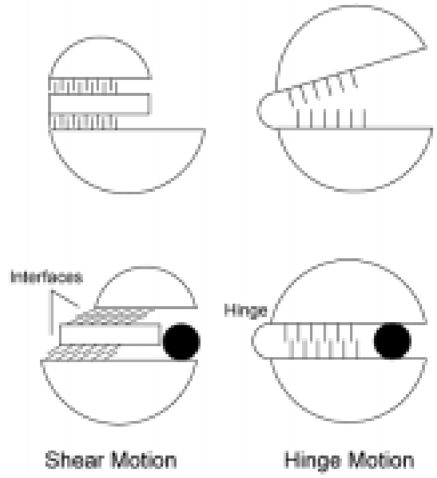
Morph Server, Hinge & Shear, Packing Based Classification

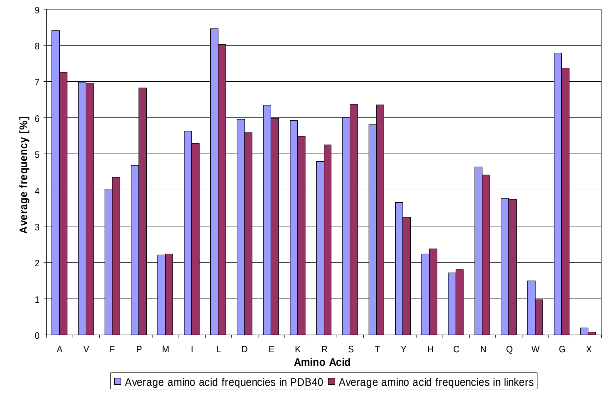
2 Analysis of Packing

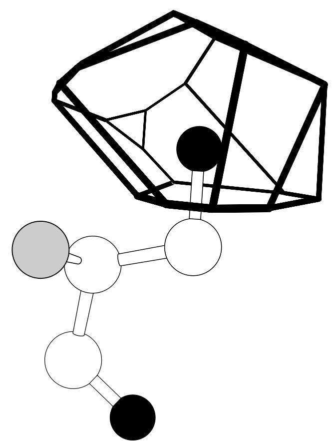
Voronoi Polyhedra, Standard Volumes

3Motion in Sequences

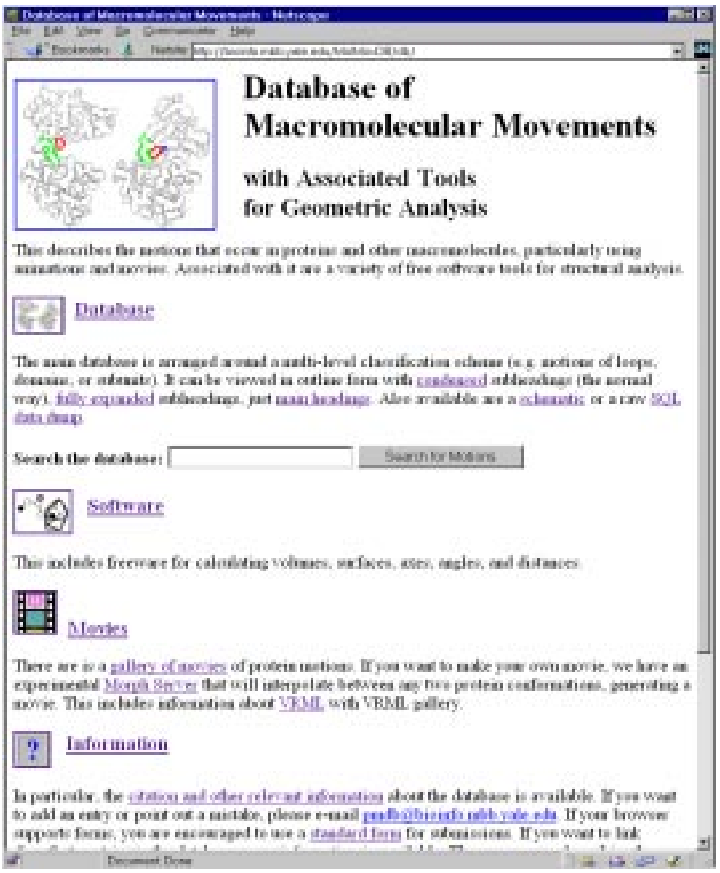
Hinge Profile, Occurrence of Mobility in Genomes



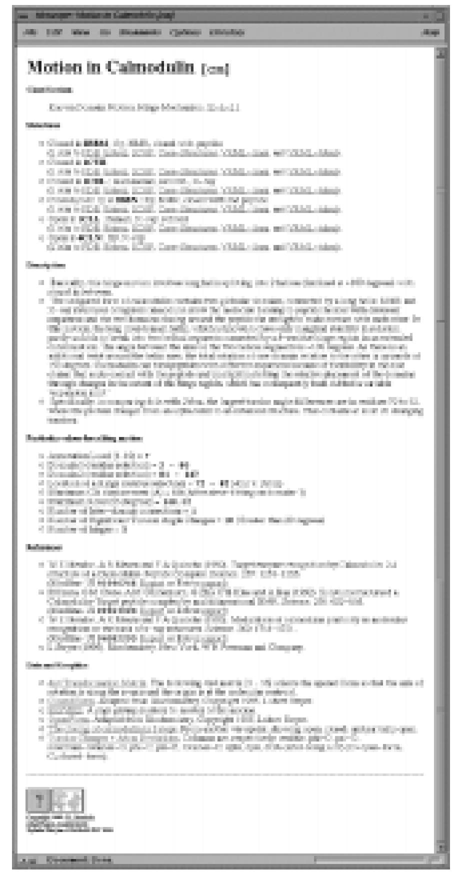




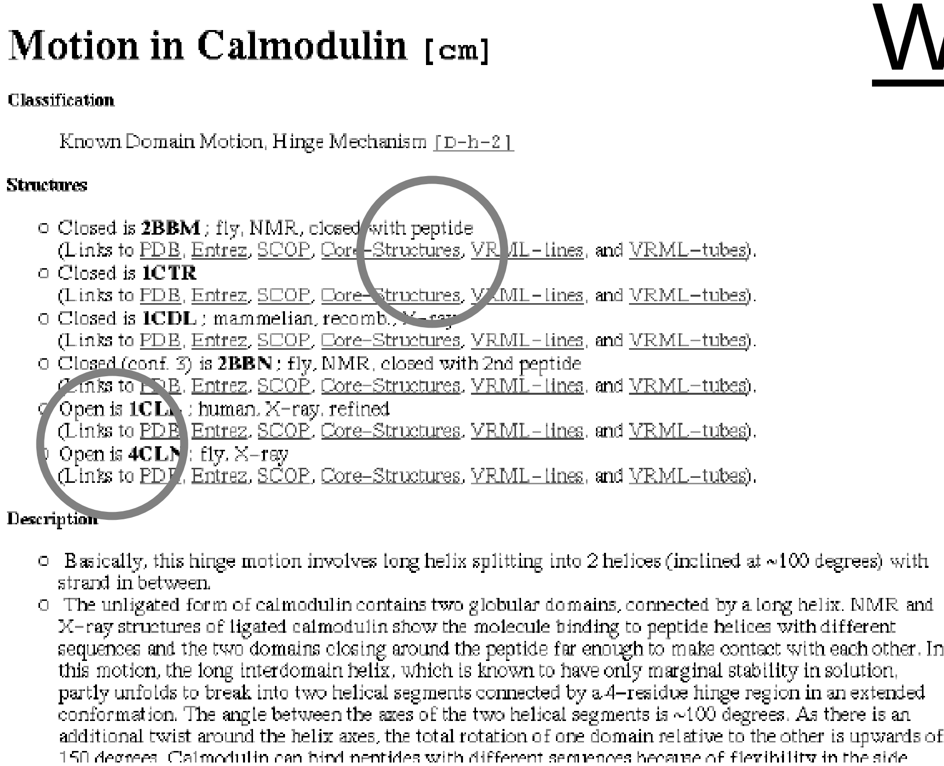
Organize Motions into a Database



Query with Calmodulin



What's in the DB? Basic



Motion in Calmodulin [cm]

Classification

Known Domain Motion, Hinge Mechanism [D-h-2]

Structures

- Closed is **2BBM** ; fly, NMR, closed with peptide
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed is **1CTR**
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed is **1CDL** ; mammelian, recomb., X-ray
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed (conf. 3) is **2BBN** ; fly, NMR, closed with 2nd peptide
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Open is **1CLL** ; human, X-ray, refined
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Open is **4CLN** ; fly, X-ray
(Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).

Description

- Basically, this hinge motion involves long helix splitting into 2 helices (inclined at ~100 degrees) with strand in between.
- The unligated form of calmodulin contains two globular domains, connected by a long helix. NMR and X-ray structures of ligated calmodulin show the molecule binding to peptide helices with different sequences and the two domains closing around the peptide far enough to make contact with each other. In this motion, the long interdomain helix, which is known to have only marginal stability in solution, partly unfolds to break into two helical segments connected by a 4-residue hinge region in an extended conformation. The angle between the axes of the two helical segments is ~100 degrees. As there is an additional twist around the helix axes, the total rotation of one domain relative to the other is upwards of 150 degrees. Calmodulin can bind peptides with different sequences because of flexibility in the side

~150 Different Motion Ids, ~250 PDB identifiers
PDB id acts as Foreign Key into other DBs
Text Blurb, Literature refs...

~20 Relational Tables

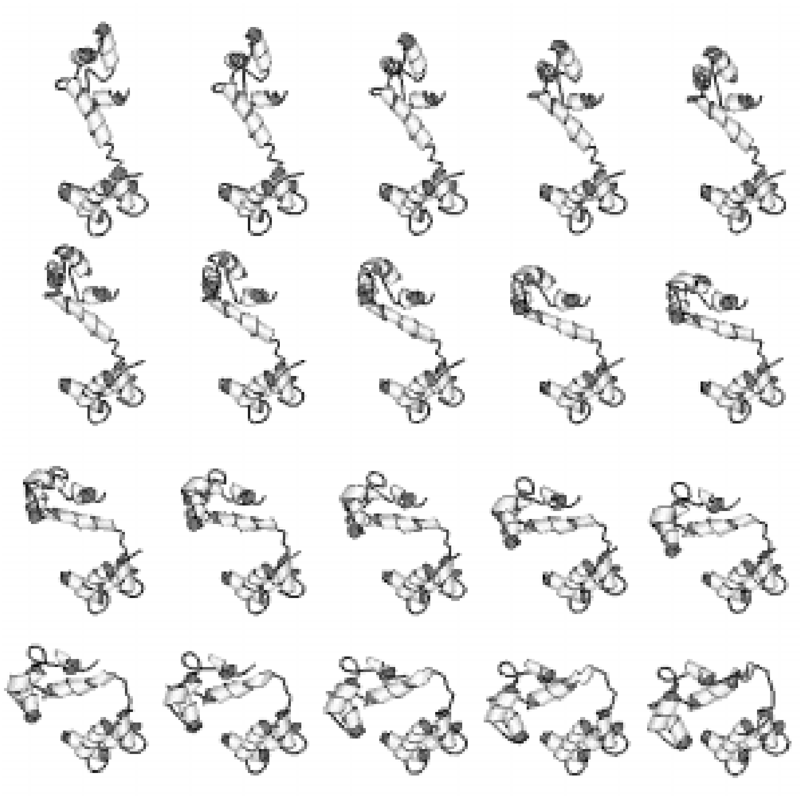
Standardized Terminology

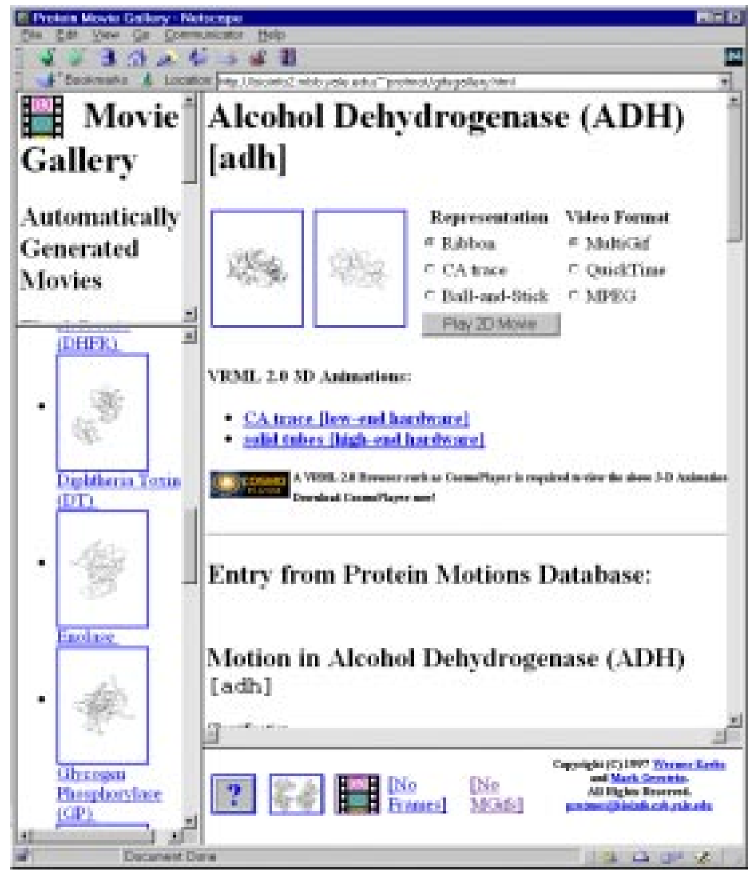
Particular values describing motion

- Annotation Level (1..10) = **7**
- Domain 1 (residue selection) = **2 - 80**
- Domain 2 (residue selection) = **81 - 147**
- Location of a Hinge (residue selection) = **72 - 82** (4cln v. 2bbm)
- Maximum CA displacement (A) = **60** (After sieve-fitting on domain-1)
- Maximum Rotation (degrees) = **148.02**
- Number of Inter-domain connections = **1**
- Number of Significant Torsion Angle Changes = **18** (Greater than 20 degrees)
- Number of hinges = **1**

- Standard statistics
  - ◊ torsion angles, max CA disp., &c
- Relations between motions
  - ◊ “sim-to”, “contains,” “Share-characteristics”
- Inferred Motions
  - ◊ 1 structure but “sim-to” another with 2

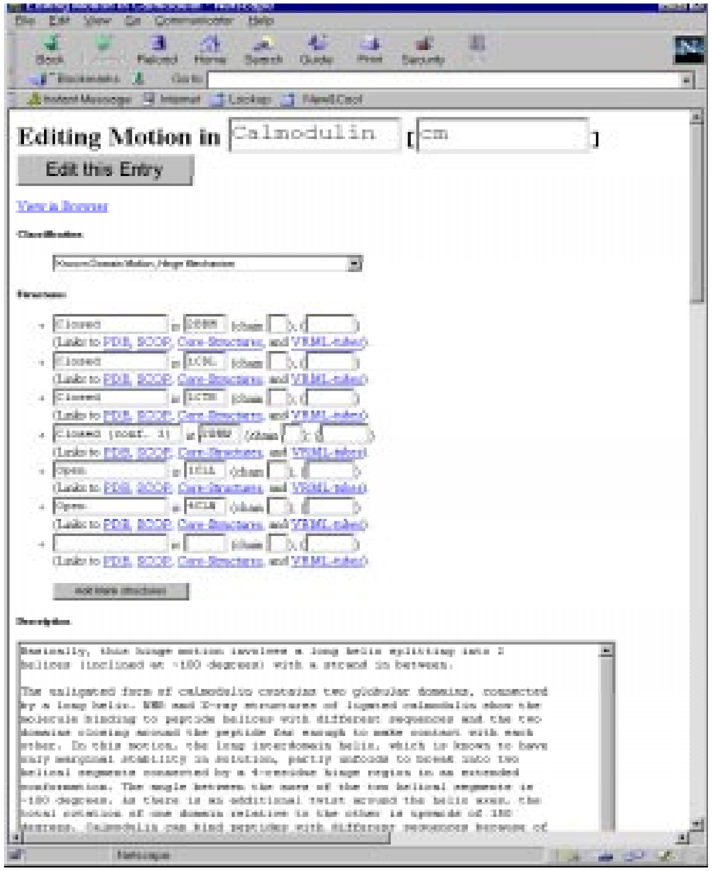
Interpolation with Packing Constraints (putting together "Morph Movies")





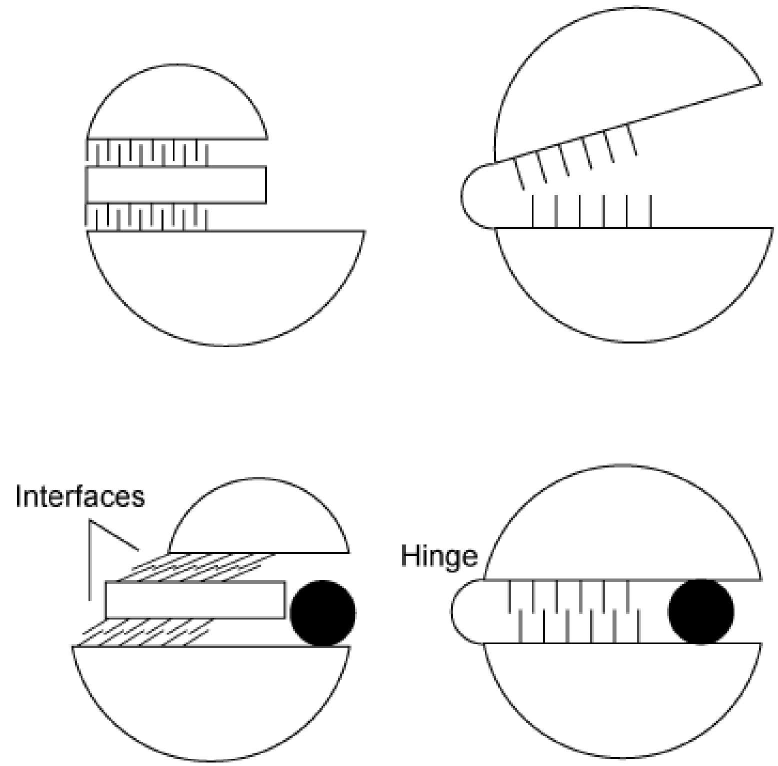
Database Issues

- IN: Manage Transactions
  - ◊ Editing by Remote Experts via forms
  - ◊ Annotation Level, Quality Control
- IN: Morph Server
  - ◊ Automatically takes 2 crystal structures and Analyzes conf. differences
  - ◊ Generates a “movie” by linear interpolation with simple restraints (bond lengths, angles, VDW interactions)
- OUT: Defined interactions with other DBs
  - ◊ Interface
- Complex Data in a Relational DB?
  - ◊ Moving to an Informix Object Relational Approach



Information, Size, then Packing Based Classification

Motion	Number Known Forms	Size of Motion	Mechanism of Motion	Examples	#
Motion	2 forms	Fragment	Hinge	TIM, LDH, TGL	14
			Shear	Insulin	3
			Unclassifiable	MS2 Coat	3
		Domain	Hinge	LF, ADK, CM	16
			Shear	CS, TrpR, AAT	8
			Refold	Serpin, RT	3
			Special	Ig elbow	1
			Unclassifiable	TBP, EF-tu	3
		Subunit	Allosteric	PFK, Hb, GP	4
			Non-allosteric	Ig VL-VH	2
			Unclassifiable		
	1 form	Fragment	Hinge		
			Shear		
			Unclassifiable	bR	1
		Domain	Refold		
			Hinge	LF~TF,SBP	10
			Shear	HK~PGK,HSP	4
			Special		
			Unclassifiable	Myosin	4
		Subunit	Allosteric		
			Non-allosteric		
			Unclassifiable	PCNA, GroEL	3

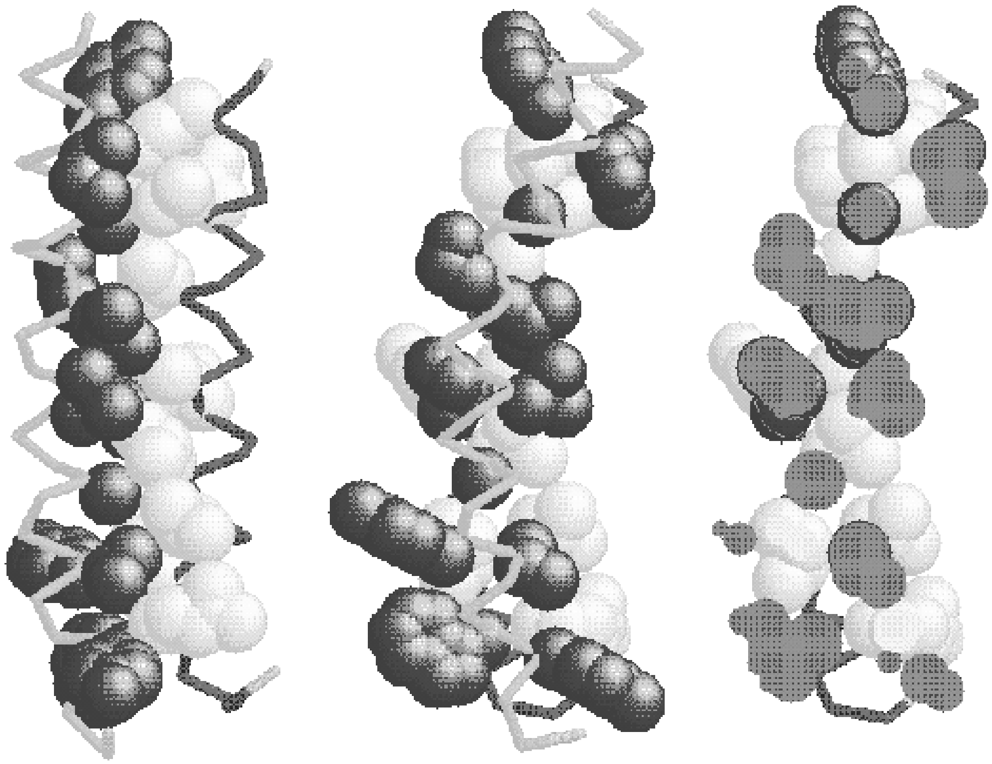


Shear Motion

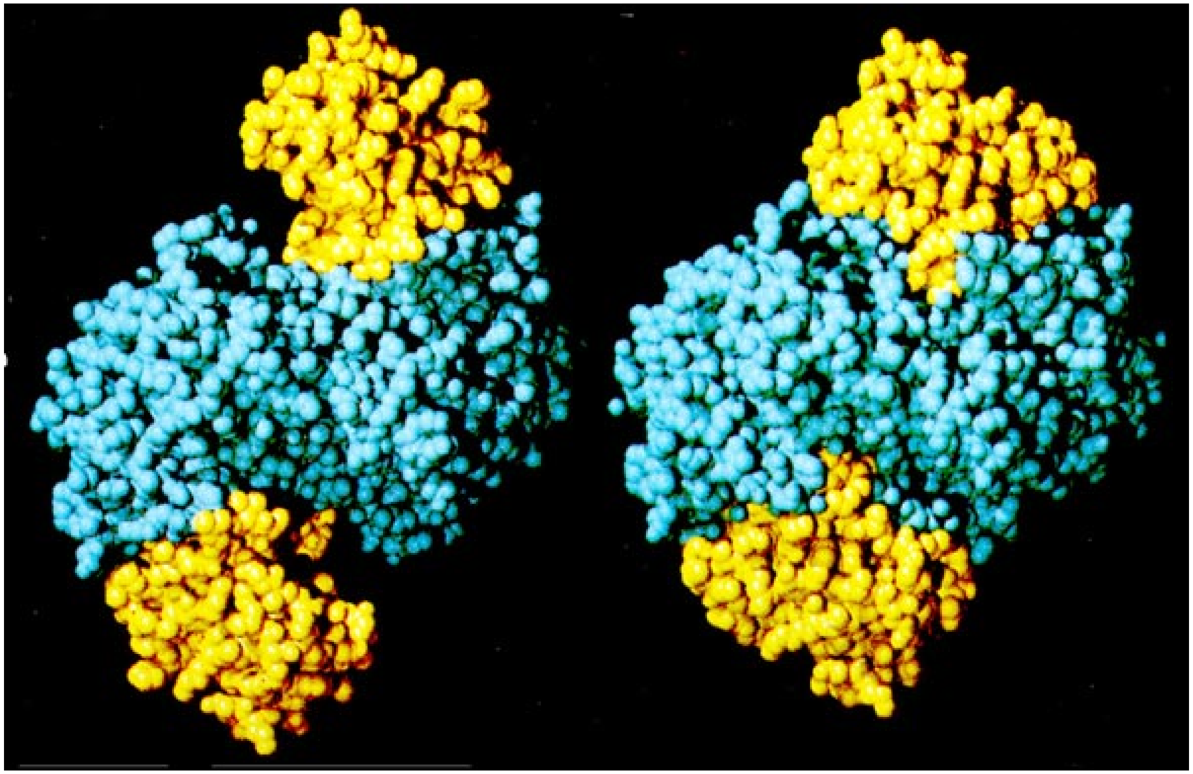
Hinge Motion

Interface Packing and Motions

- Intercalcating Interface, Knobs into Holes
- Packing is a strong constraint on motions
  - ◊ Domain or loop motions have to be fast (~10 ps – 100 ns)
  - ◊ Can't cross big energy barriers involved in repacking an interface
- Not applicable to allosteric motions, which are much slower (~1 ms) and do involve repacking interfaces

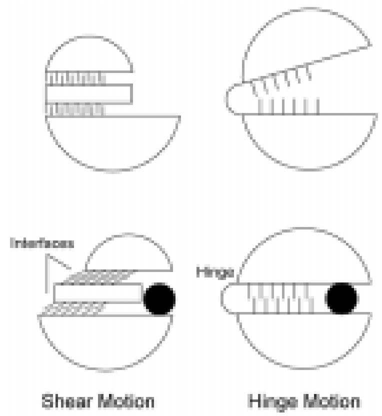


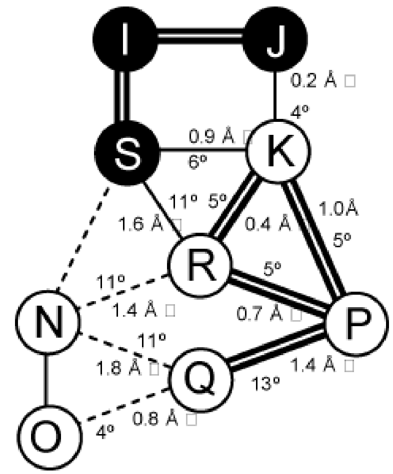
Motion in Citrate Synthase



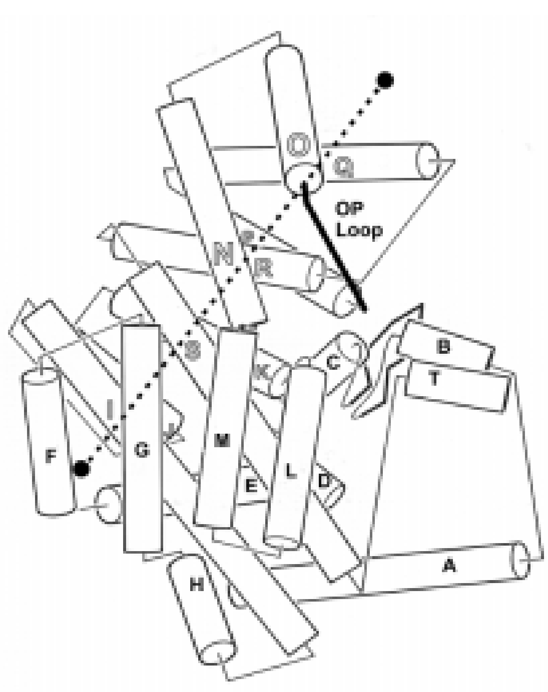
Packing Based Classification:

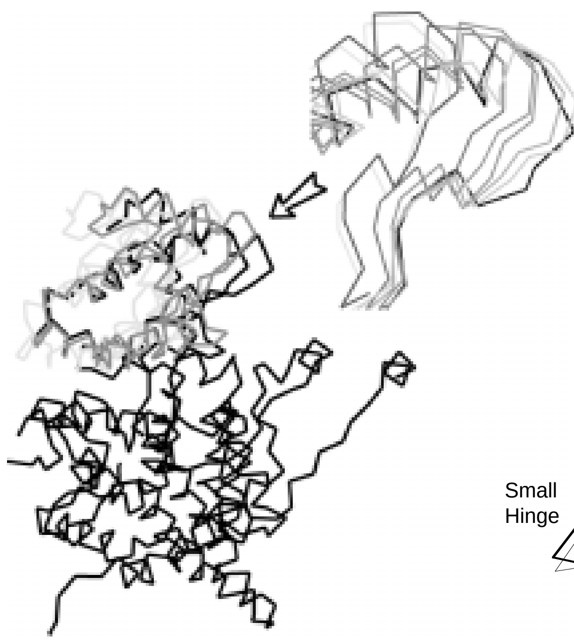
Hinge v ***Shear***

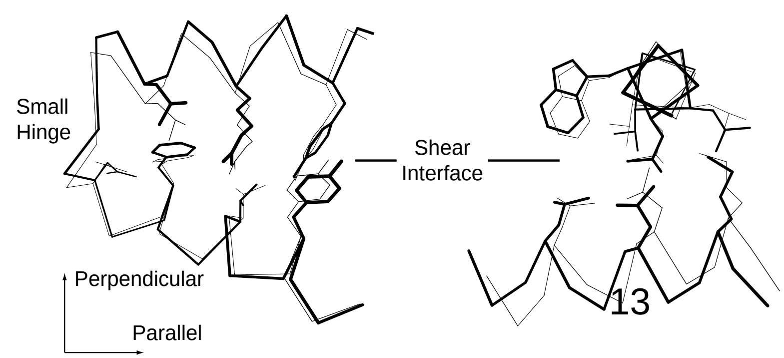




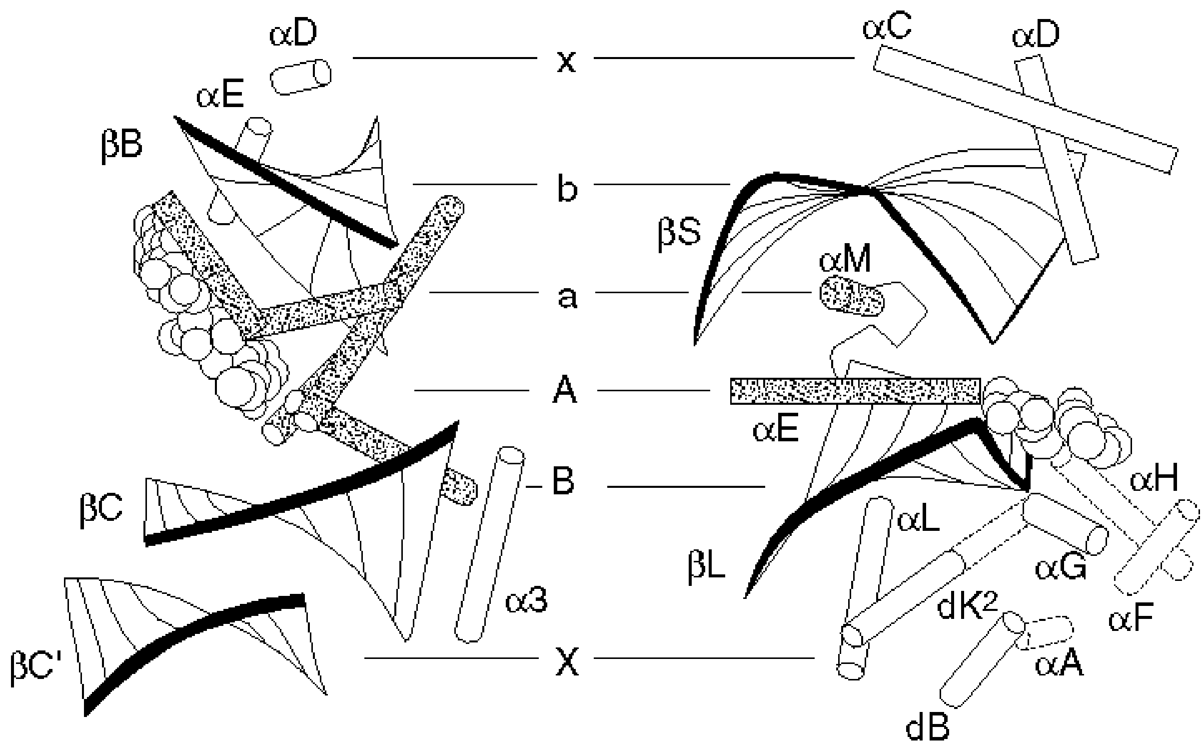
Shear Mechanism Involves Many Small Motions across a <u>Continuously Maintained Interface</u>







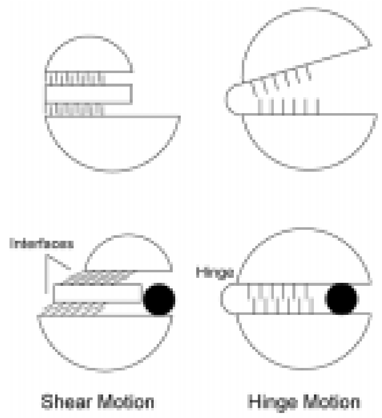
Proteins With Shear Motions are Often Divided into Layers



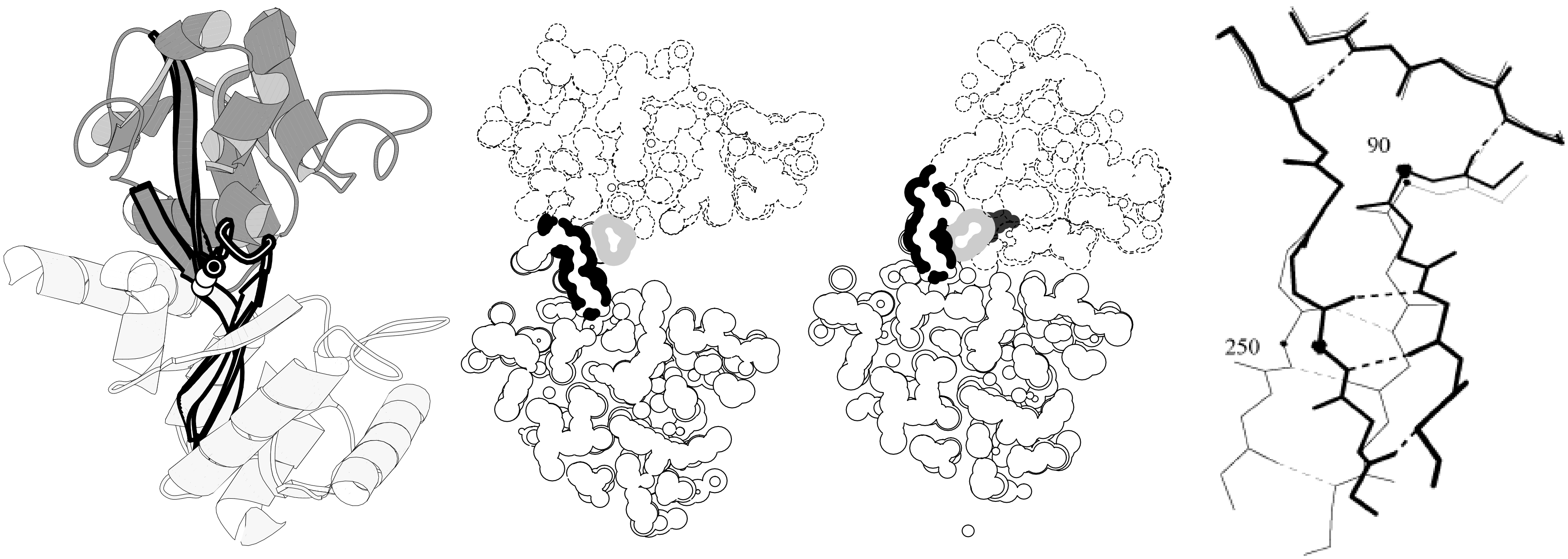
GAPDH

Hexokinase

Packing Based Classification: ***Hinge*** v Shear

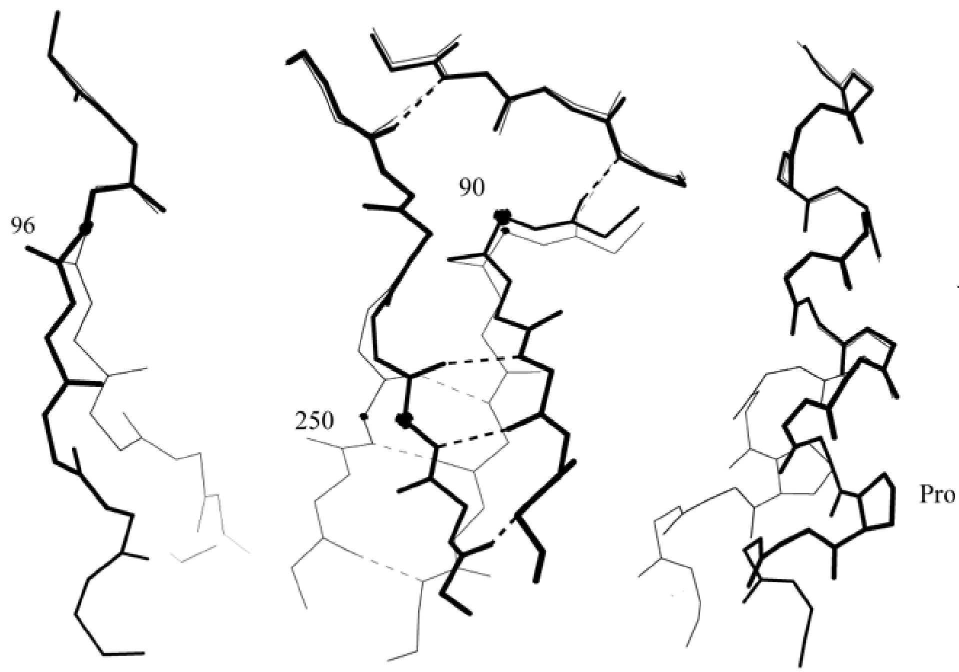


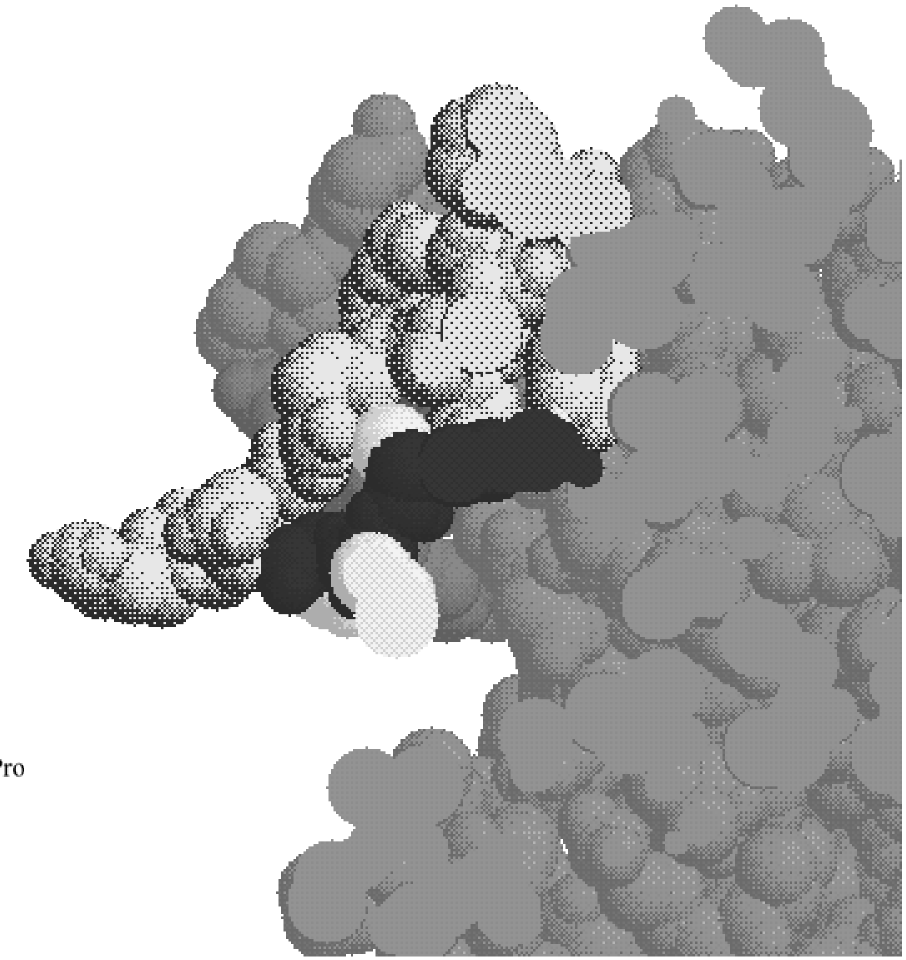
Hinge Mechanism involves absence of steric constraints (continuously maintained interface), esp. at hinge



Absence of Tight Packing at Hinge

Chain Topology is not important





Studying Macromolecular Motions in a Database Framework: from Structure to Sequence

1Motions Database

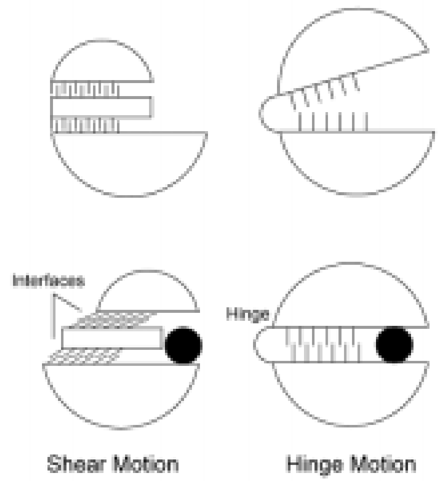
Morph Server, Hinge & Shear, Packing Based Classification

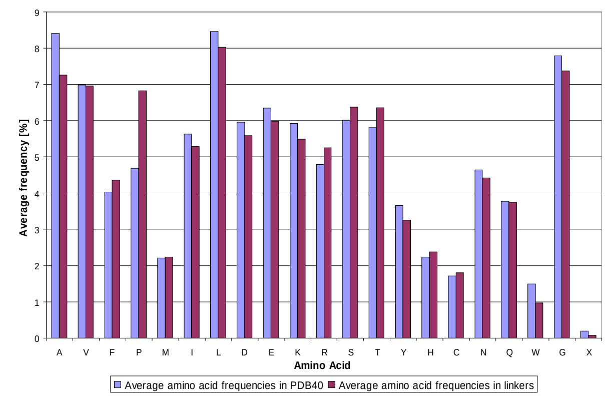
2 Analysis of Packing

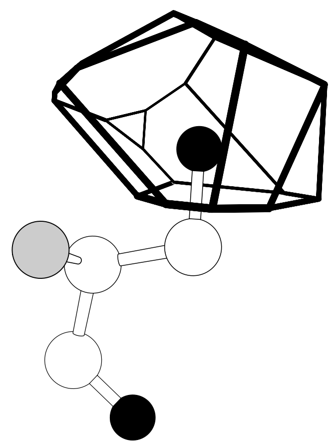
Voronoi Polyhedra, Standard Volumes

3Motion in Sequences

Hinge Profile, Occurrence of Mobility in Genomes

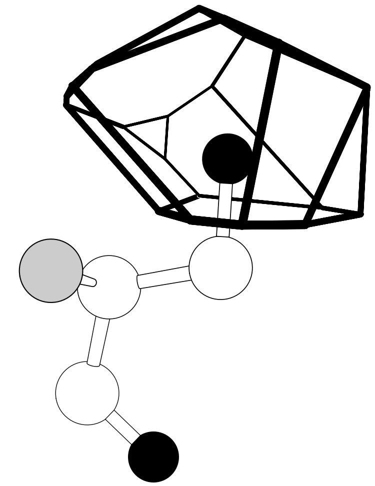


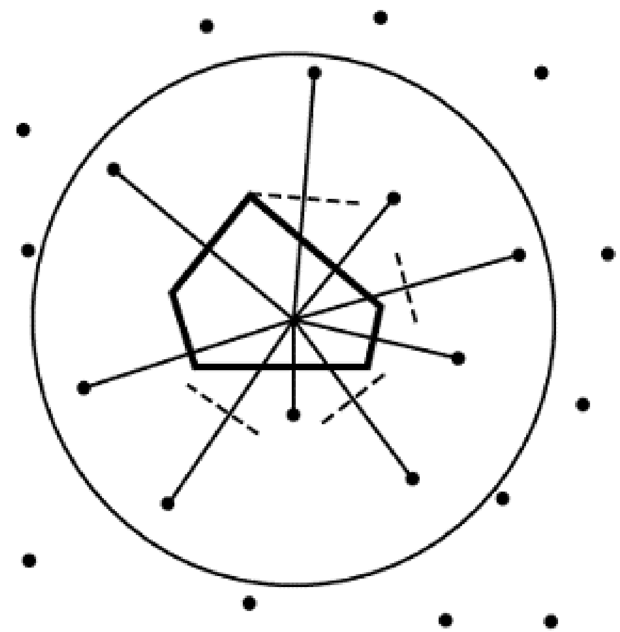




Quantify Packing and Contacts with Voronoi Polyhedra

- Each atom surrounded by a single convex polyhedron and allocated space within it
  - ◊ Allocation of all space (large V implies cavities)
- 2 methods of determination
  - ◊ Find planes separating atoms, intersection of these is polyhedron
  - ◊ Locate vertices, which are equidistant from 4 atoms





Voronoi Volumes, the Natural Way to Measure Packing

Packing Efficiency

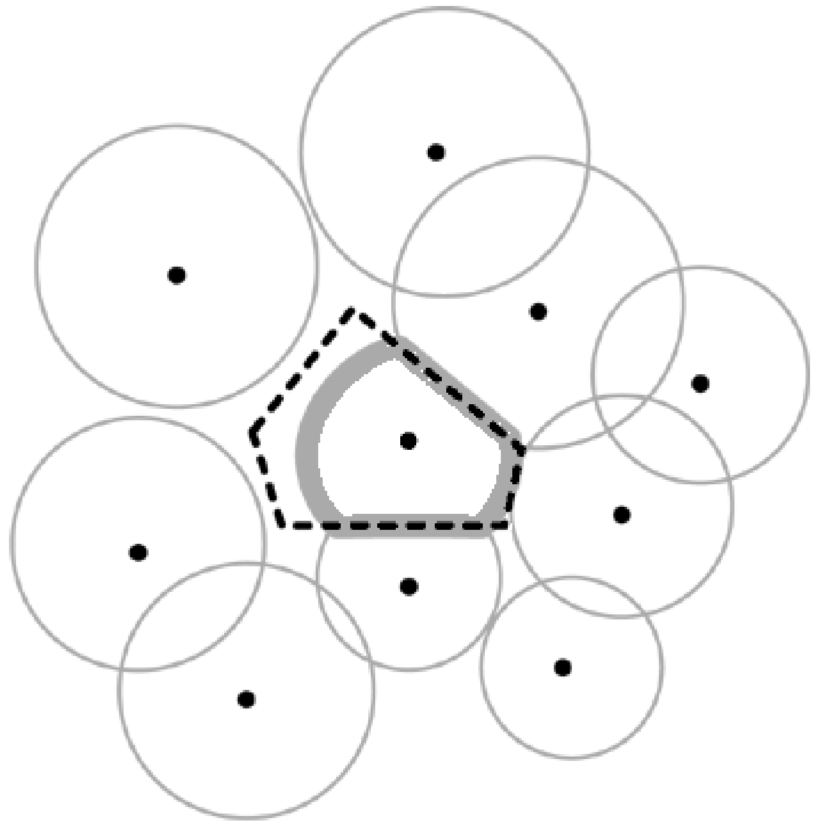
```
= Volume-of-Object
  -----------------
  Space-it-occupies
```

= V(VDW) / V(Voronoi)

- Absolute v relative eff.

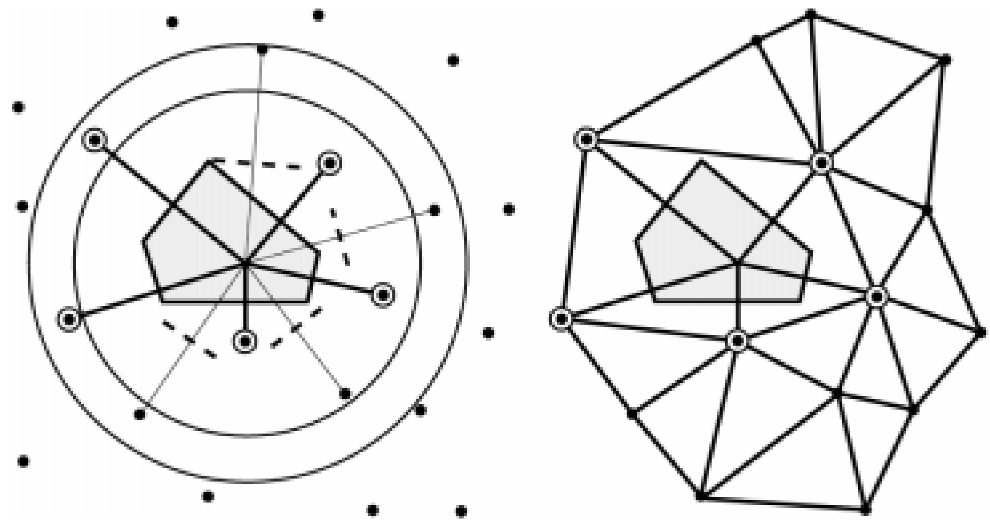
  V1 / V2

- Other methods
  - ◊ Measure Cavity Volume (grids, constructions, &c)



Delauney Triangulation, the Natural Way to Define Packing Neighbors

- Related to Voronoi polyhedra (dual)
- What "coordination number" does an atom have? Doesn't depend on distance
- alpha shape



Compare with Std. Volumes to Quantify Packing

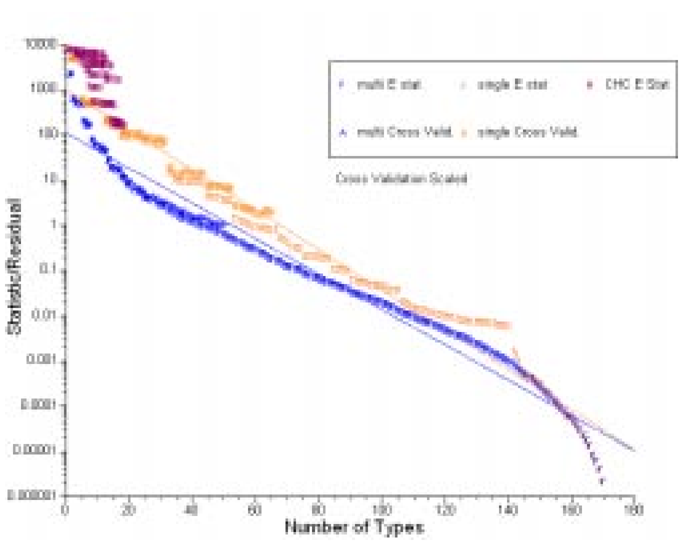
- Standard Volumes for atoms and residues in the buried core of proteins
- Measure Packing at Interfaces by comparing the volumes of atom with these standard

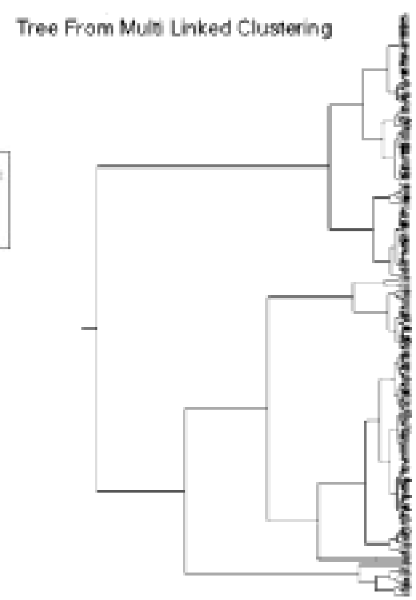
G	64	c	105	T	120	V	139	H	159	M	168	R	194
A	90	C	113	P	124	E	140	L	165	K	170	Y	198
S	94	D	117	N	128	N	150	I	165	F	193	W	233

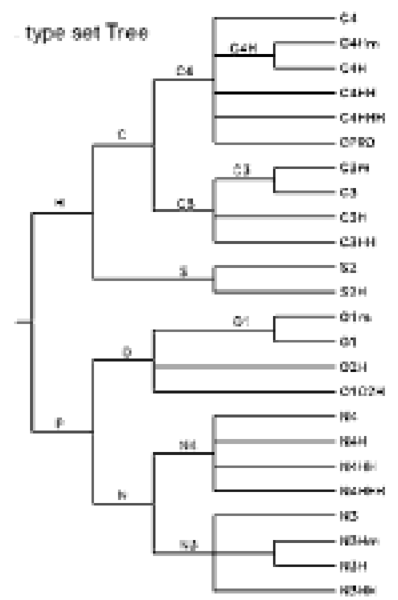
mainchain vol.			non-polar sidechain vol.			polar sidechain vol.		
atom	core	surf.	atom	core	surf.	atom	core	surf.
N	14.0	14.8	>CH-	14.7	15.2	–NH2	23.4	24.8
CA	13.5	14.2	-CH2-	23.7	24.5	–OH	17.3	17.8
C	9.3	10.0	–CH3	36.6	37.6	=O	16.8	18.3
O	15.9	16.6	>C=	10.1	10.6	–O (-)	16.0	16.7

Clustering into a set of Atom Types

- Which atoms are equivalent? How many types valid?
- 18 types, [CNOS][34]H[123][bsu]

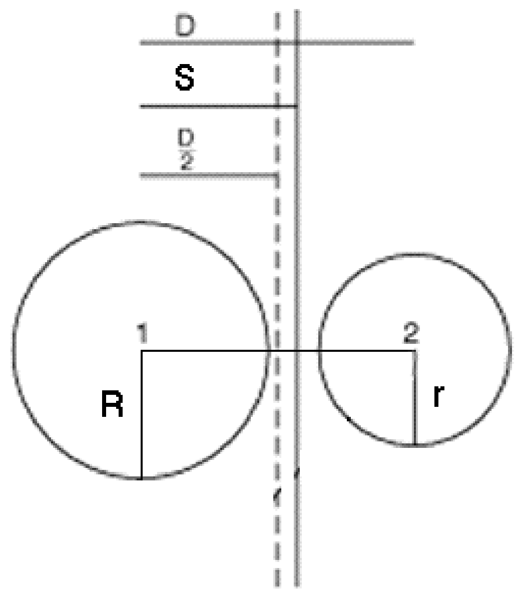






Atoms have different sizes

- Difficulty with Voronoi Meth. Not all atoms created equal
- Solutions
  - ◊ Bisection -- plane midway between atoms
  - ◊ Method B (Richards) Positions the dividing plane according to ratio
  - ◊ Radical Plane
- VDW Radii Set



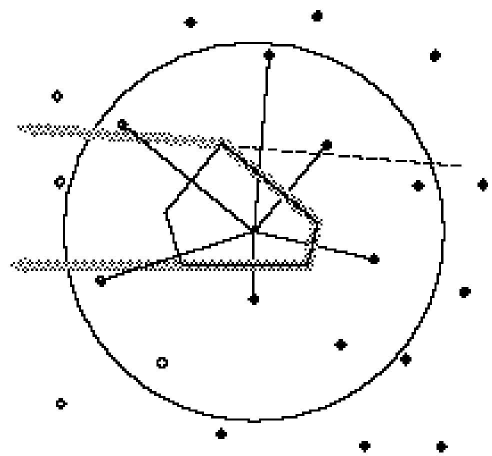
Set of VDW Radii

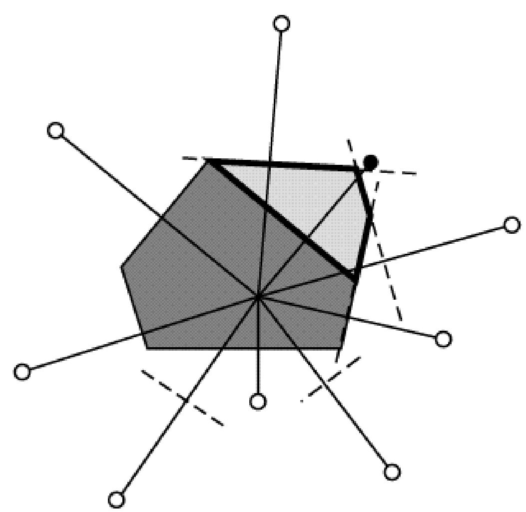
- Great differences in a sensitive parameter (Radii for carbon 1.87 vs 2.00)
- Complex calculation: minimizing SD, iterative procedure, from protein structures
- Look for common distances in CCD
- Preliminary Solution

Atom	Bondi	New
C4__	1.87	1.88
C3H1	1.76	1.76
C3H0	1.76	1.61
O1H0	1.40	1.42
O2H1	1.40	1.46
N___	1.65	1.64
S___	1.85	1.77

Other Aspects of Calculation of Standard Volumes

- Problem of the Protein Surface: Atom Selection, Definition of Buried Atoms
- Collection of a Standard Set of Hi-resolution Structures (via scop)
- Different Algorithm of Polyhedra Construction (Chopping Down) Avoids symmetry center problem



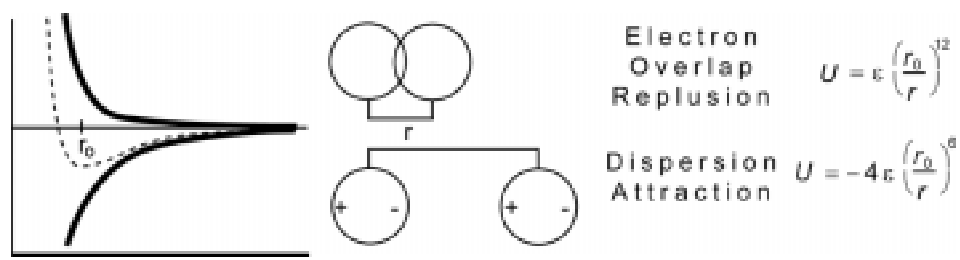


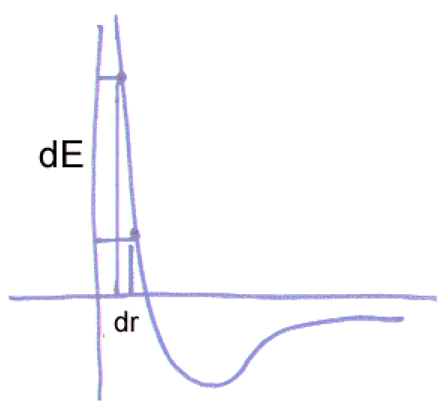
Sample Results of Volume Calculations: Significant but Small Changes in Packing

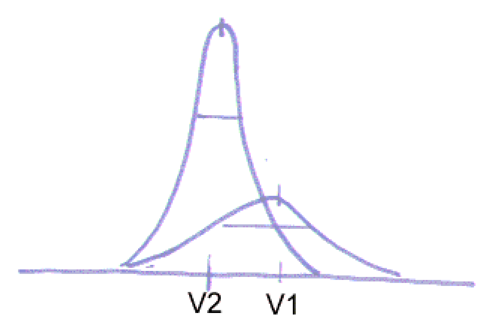
-CH2- volume in cubic A

23.7	**standard volume in protein core**
23.6	**mobile helix-helix interface (CS, TrpR)**
24.8	**grooves on protein surface (in high-res. struc. via SurFractal)**

- VDW ~ Packing
  - ◊ Exponential Repulsion
- Many observations (>10K) in standard volumes gives small error about the mean (SD/sqrt(N))







Studying Macromolecular Motions in a Database Framework: from Structure to Sequence

1Motions Database

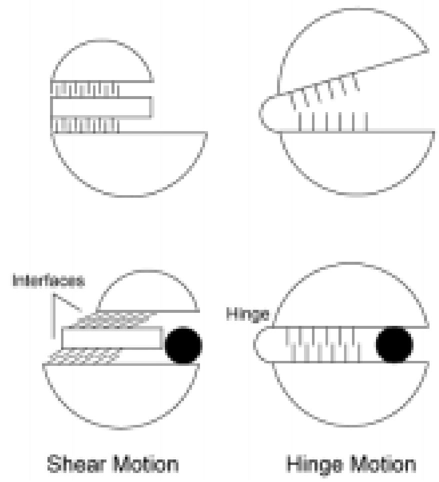
Morph Server, Hinge & Shear, Packing Based Classification

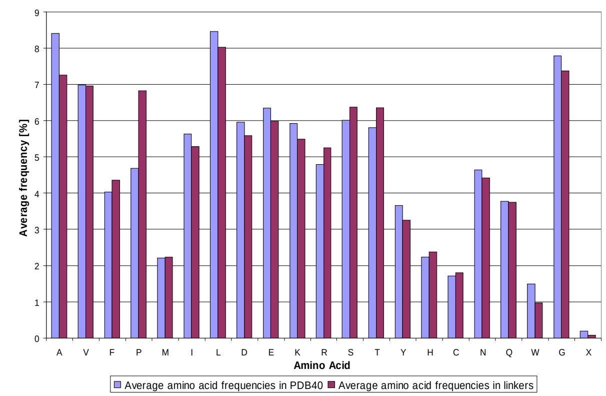
2 Analysis of Packing

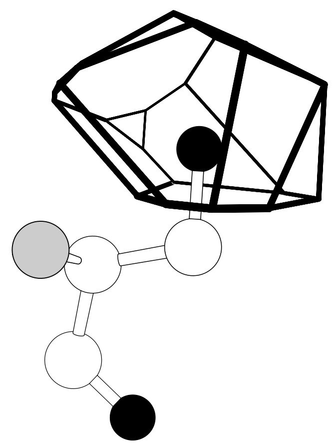
Voronoi Polyhedra, Standard Volumes

3Motion in Sequences

Hinge Profile, Occurrence of Mobility in Genomes





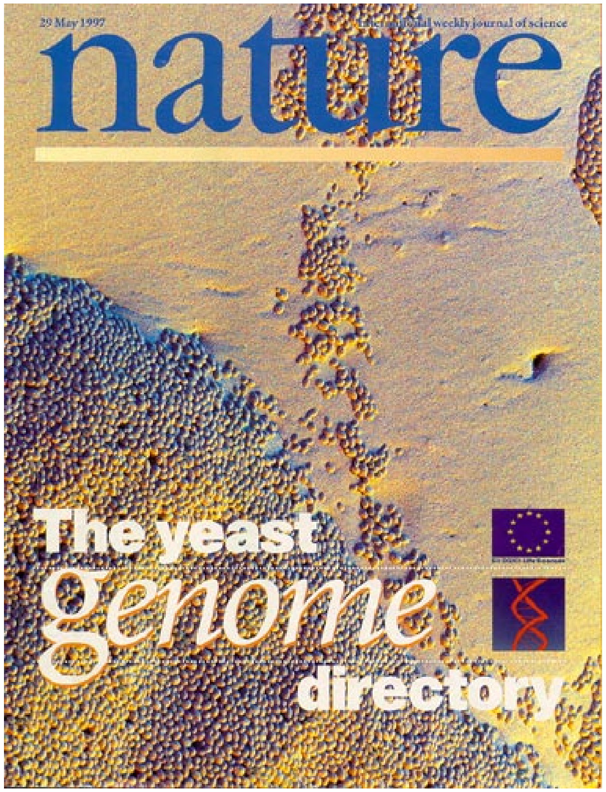


Genomes highlight the **Finiteness** of Biology

1995

Bacteria
1.6 Mb, ~1600 genes
[Fleischmann *et al.* (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd." *Science* **269**: 496-512.]

1997



Eukaryote
13 Mb, ~6000 genes

1998.....

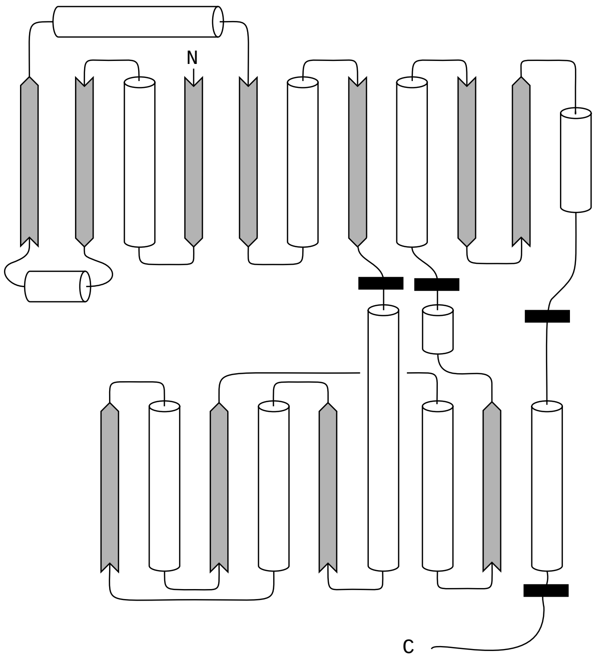
Microbial Genomes
>15 completed,
~40 underway

The Worm:
75% of 100 Mb
done, with ~13 K
genes so far)

The Human:
3 Gb & 100 K
genes, 2003?...

Extrapolate Structural Information to Genome Sequences

- Structures are the "gold standard"
  - ◊ Gives relation to chemistry
- Focus on defining Interdomain Linker, both flexible hinges and rigid connections
- What to know:
  - ◊ What are the sequence characteristics of a mobile region?
  - ◊ How many mobile regions in a genome?



Linker ID	Linker Consensus Sequence
4cln	MARKMKDTDSE
6ldh	AGARQQEGESRLNLVQRNVNIFKF
adenkin1	VPFEVI
adenkin2	LRLTA
adenkin3	GEPLIQRDDDKE
adenkin4	AYHAQTE
anxbreat	MKGAGT
anxtrp1	YEAGELKWG
anxtrp2	EETIDRET
dt	LFQVVHNS
enolase	GASTGIY
enolase2	SDKS
lfh_hinge1	QTHY
lfh_hinge2	RVPS
ras	AGQEEYSAMRDQYMR
tbsv	PQPTNTL

2. Find Sequence Homologues (FASTA), Multi-alignment

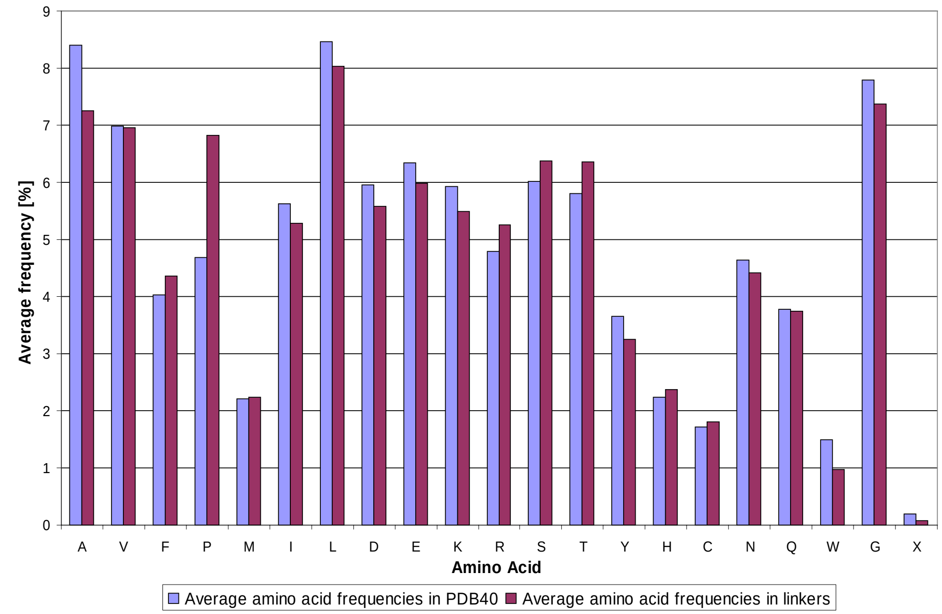
OWL ID	Sequence
CALN_CHICK	MARKMKDTDSE
MUSCAMC	MARKMKDTDSE
CALM_PATSP	MARKMKDTDSE
CALM_PYUSP	MARKMKDTDSE
CALM_METSE	MARKMKDTDSE
CALM_STIJA	MARKMKDTDSE
CALM_HUMAN	MARKMKDTDSE
CALM_DROME	MARKMKDTDSE
HSCAM3X1	MARKMKDTDSE
CALM_EMENI	MARKMKDTDSE
CALM_NEUCR	MARKMKDTDSE
CALM_ELEEL	MAKKMKDTDSE
NEUCLMDLN	MARKMKDTDSE
SSO4B01	MARKMKDTDSE
CALL_ARBPU	MARKMKETDSE
CALM_PLECO	MARKMRDTDSE
CALL_HUMAN	MARKMKDTDNE
CALS_CHICK	MARKMRDSDSE
CALM_PHYIN	MARKMKDTDSE
CALM_PNECA	MARKMKDVDSE
CALM_TRYBB	MARKMQDSDSE
CALM_TRYCR	MARKMQDSDSE
S53019	MARKMKDTDSE
TRBCMRSG	MARKMQDSDSE
CALM_HORVU	MARKMKDTDSE
JC1033	MARKMKDTDSE
CAL1_PETHY	MARKMKDTDSE
CAL6_ARATH	MARKMKDTDSE

1. Extract from Structures (i.e. Motions DB) few "Gold Standard" hinges

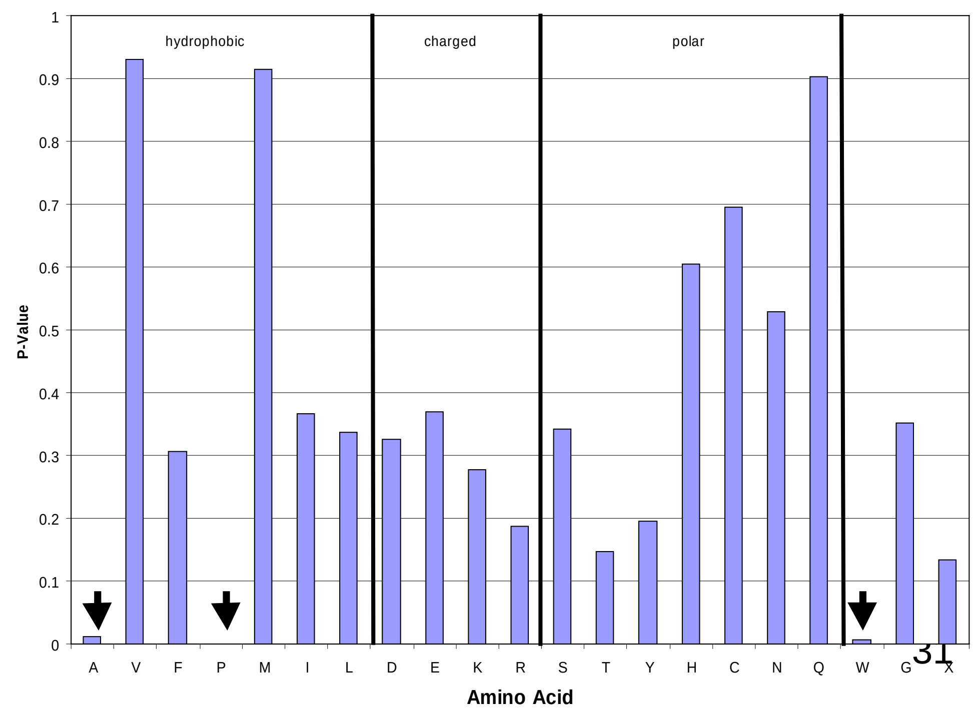
Residue	Propensity
A	1.3268
C	0.1097
D	1.1684
E	1.4702
F	0.5624
G	1.2972
H	0.4806
I	0.4462
K	1.0519
L	0.5303
M	2.6603
N	0.7729
P	0.4051
Q	1.8076
R	1.8013
S	0.8269
T	0.9002
V	0.6865
W	0.308
Y	1.3375

3. Build a scale for residues to occur in flexible hinges

Significance Values for Compositional Differences



Compare Observed Difference in Linker Composition to that Found in PDB (via Resampling) to get a P-value
(P up, A & W down)



Multi-position Patterns

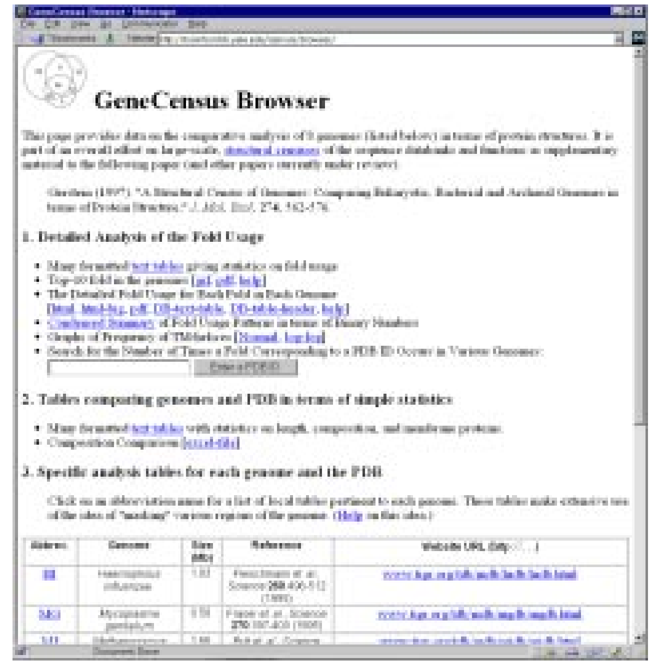
Profile or HMM of a Linker Region

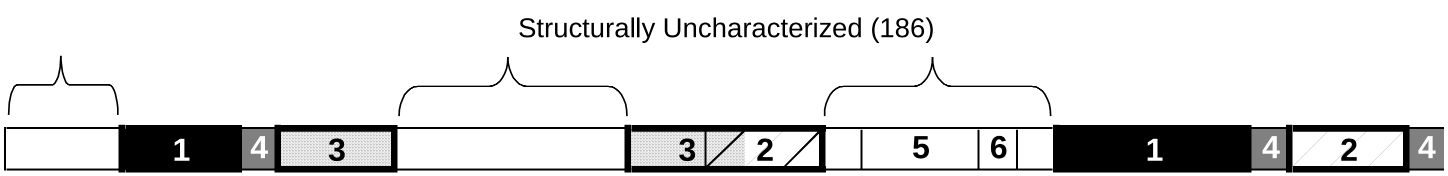
																	PDB40 average
A	8.6	7.8	4.7	5.6	6.0	8.6	9.5	5.6	4.7	6.5	5.6	7.3	6.9	9.1	9.5	9.9	8.4
V	6.0	8.2	8.2	6.0	8.2	5.6	9.1	6.0	8.2	4.7	6.0	4.7	7.3	9.1	5.2	8.6	7.0
F	4.7	3.9	6.5	3.5	2.6	2.6	6.0	2.6	4.7	3.0	4.3	6.0	5.2	4.3	4.3	5.6	4.0
P	3.9	6.5	6.0	6.0	5.2	9.1	6.9	10.8	9.1	10.3	9.9	6.0	8.6	2.6	4.7	3.5	4.7
M	4.7	1.3	1.3	2.6	2.6	0.0	1.7	1.7	4.3	3.0	1.3	1.3	2.2	1.7	3.0	3.0	2.2
I	5.6	3.5	7.3	6.5	3.9	6.0	3.9	3.5	5.2	6.9	4.7	2.6	4.7	8.6	5.6	6.0	5.6
L	11.6	9.1	11.2	6.0	16.4	7.3	4.3	6.5	8.2	3.5	7.3	5.2	7.3	6.5	10.3	7.8	8.5
D	4.7	6.5	6.0	3.9	6.0	4.7	5.6	8.6	4.3	3.9	3.5	7.3	6.9	7.3	4.3	5.6	6.0
E	5.2	5.2	3.9	6.5	4.7	4.7	7.8	4.7	6.5	4.3	6.5	9.1	7.3	5.2	8.6	5.6	6.3
K	5.2	6.5	3.9	5.6	5.2	6.9	4.7	4.7	6.0	7.8	3.9	6.5	5.2	5.2	3.0	7.8	5.9
R	5.2	3.9	4.7	9.1	6.5	5.2	5.2	5.6	5.6	4.7	6.0	5.2	5.2	4.7	3.0	4.3	4.8
S	7.8	6.0	5.2	6.9	6.5	8.2	6.9	6.5	3.5	6.0	9.5	7.8	4.3	3.9	8.6	4.7	6.0
T	4.7	5.6	3.0	5.6	6.5	9.5	6.9	6.0	6.5	11.2	7.3	6.5	6.0	4.7	8.2	3.5	5.8
Y	2.2	3.9	6.5	3.0	3.5	2.2	2.6	3.5	2.2	3.9	2.6	2.2	3.0	3.5	3.5	4.3	3.7
H	1.7	3.5	3.0	3.5	3.5	2.6	3.5	2.2	2.2	0.9	1.7	2.2	1.7	2.6	1.3	2.2	2.2
C	1.7	2.6	0.9	1.3	1.7	2.6	0.4	2.2	0.9	1.3	4.7	1.7	1.7	3.9	0.4	0.9	1.7
N	4.7	3.9	3.5	6.5	3.0	4.3	2.6	3.0	5.6	5.2	3.5	6.5	3.9	6.0	3.0	5.6	4.6
Q	3.9	5.2	3.5	5.2	2.6	0.9	3.0	2.2	3.5	4.7	3.5	2.2	6.5	4.3	4.3	4.7	3.8
W	1.3	0.9	0.9	2.6	0.4	0.9	0.4	0.9	0.4	1.3	0.0	1.3	0.4	0.9	2.2	0.9	1.5
G	6.0	6.0	9.9	4.3	5.2	8.2	9.1	13.4	8.2	6.9	8.2	8.6	5.6	6.0	6.9	5.6	7.8
X	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	

A	.908	.728	4e-2	.125	.196	.908	.562	.125	4e-2	.293	.125	.561	.415	.729	.562	.416	hydrophobic
V	.577	.481	.481	.577	.481	.417	.224	.577	.481	.184	.577	.184	.841	.224	.285	.338	
F	.598	.911	.059	.666	.276	.276	.126	.276	.598	.449	.836	.126	.393	.836	.836	.235	
P	.573	.207	.346	.346	.737	2e-3	.114	5e-5	2e-3	1e-4	3e-4	.346	4e-3	.134	.971	.385	
M	1e-2	.366	.366	.717	.717	2e-2	.637	.637	3e-2	.433	.366	.366	.961	.637	.433	.433	
I	.990	.155	.267	.585	.257	.793	.257	.155	.772	.408	.571	4e-2	.571	5e-2	.990	.793	
L	.084	.754	.136	.186	3e-5	.541	2e-2	.280	.882	6e-3	.541	.071	.541	.280	.312	.705	
D	.442	.750	.966	.185	.966	.442	.821	.089	.296	.185	.108	.389	.556	.389	.296	.821	charged
E	.476	.476	.127	.936	.327	.327	.384	.327	.936	.211	.936	.092	.545	.476	.158	.653	
K	.638	.730	.194	.842	.638	.538	.457	.457	.945	.243	.194	.730	.638	.638	.061	.243	
R	.793	.530	.974	2e-3	.240	.793	.793	.575	.575	.974	.389	.793	.793	.974	.215	.742	
S	.269	.990	.599	.578	.774	.166	.578	.774	.101	.990	2e-2	.269	.283	.176	.095	.425	polar
T	.498	.897	.069	.897	.673	2e-2	.485	.886	.673	5e-4	.328	.673	.886	.498	.121	.127	
Y	.234	.864	2e-2	.619	.872	.234	.402	.872	.234	.864	.402	.234	.619	.872	.872	.612	
H	.619	.237	.455	.237	.237	.740	.237	.939	.939	.166	.619	.939	.619	.740	.354	.939	
C	.997	.336	.345	.647	.997	.336	.139	.634	.345	.647	2e-2	.997	.997	2e-2	.139	.345	
N	.942	.597	.404	.193	.251	.820	.143	.251	.500	.710	.404	.193	.597	.326	.251	.500	
Q	.937	.281	.804	.281	.359	2e-2	.562	.206	.804	.460	.804	.206	3e-2	.684	.684	.460	
W	.810	.459	.459	.193	.197	.459	.197	.459	.197	.810	.055	.810	.197	.459	.452	.459	
G	.324	.324	.233	5e-2	.139	.823	.482	1e-3	.823	.621	.823	.643	.218	.324	.621	.218	
X	.717	.717	.752	.752	.752	.752	.752	.752	.717	.752	.752	.752	.752	.752	.752	.752	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	

Looking for Mobile Proteins in Genomes with GeneCensus System

- GeneCensus System to Identify Structural Regions (structure match, TM-helix, &c)
- Recent Microbial Genomes: HI, MG, MJ, SC, SS, HP, EC, TP, MP
- Linkers = <50 residues between known domains or C- or N- terminal extension
- Low Complexity Region = Repetitive Sequence (AAGAAGAAAG, TSVVVTSVVVVVTSVVVTS)





1 PDB Match (152)
2 Low Complexity Region (116)
3 TM helix (30)
4 Linker Region (5)
5 Coiled-Coil
6 All-alpha or All-beta Region

Occurence of Highly Mobile Regions in the Genome

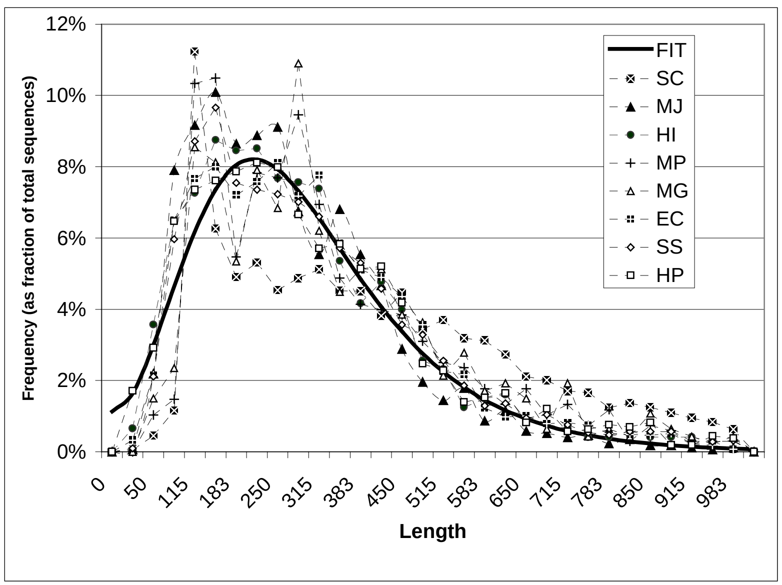
- Linkers (interdomain + extended C and N-term.) occupy ~5% of the genome and are in ~50% of genes
- Low-Complexity Regions occupy ~20% of the genome and are in ~40% of genes

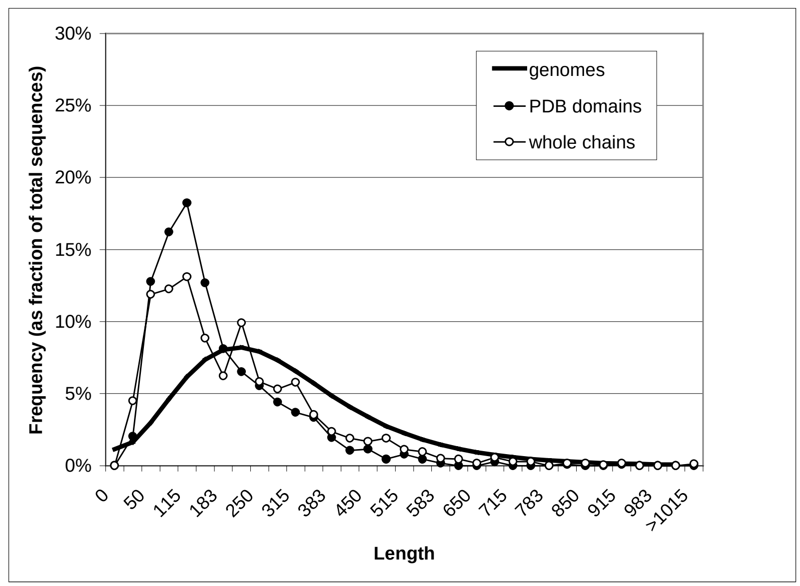
	AVG	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Statistics for Amino Acids										
Total Number	775998		1358465	505279	500616	170400	497968	237905	2900670	1033450
Fraction Masked by...										
PDB Match	8.7%	3.7%	11.1%	13.7%	8.8%	12.9%	7.1%	9.7%	6.2%	9.0%
Low-Complexity Region	21.7%	6.9%	16.7%	13.9%	22.2%	28.2%	35.1%	24.7%	23.9%	20.5%
TM-helix	4.9%	1.4%	7.3%	6.1%	4.8%	3.8%	2.9%	4.5%	5.2%	5.9%
Linker Region	5.1%	0.4%	5.3%	4.8%	4.8%	5.0%	5.0%	5.2%	4.6%	5.1%
Fraction Remaining										
Uncharacterized	59.7%	8.9%	59.6%	61.5%	59.4%	50.2%	49.9%	55.8%	60.0%	59.6%
Statistics for ORFs										
Total Number	2206	1731	4290	1680	1577	468	1735	677	6218	3168
Fraction Containing...										
PDB Match	12.6%	4.8%	14.1%	16.8%	12.2%	19.2%	11.0%	14.2%	13.5%	13.2%
Low-Complexity Region	43.0%	12.6%	34.6%	30.6%	43.2%	51.7%	61.3%	49.3%	56.3%	39.6%
TM-helix	28.8%	6.6%	34.6%	27.7%	26.9%	26.7%	19.6%	28.1%	35.6%	36.8%
Linker Region	51.0%	9.1%	49.0%	46.1%	50.4%	58.8%	55.0%	56.0%	57.3%	52.8%
Fraction Containing...										
Uncharacterized Region	76.8%	4.4%	75.2%	73.2%	75.4%	74.8%	68.8%	77.8%	84.0%	79.4%
Characterized Region	65.5%	13.7%	64.2%	58.6%	65.2%	74.1%	74.9%	70.9%	79.1%	68.3%

Simple Statistics: Distribution of Sequence Lengths

Genomes Sequences are Significantly longer than those in Known Structures

340 aa for avg. genome seq.
(470 aa for yeast)
205 aa for PDB chain
170 aa for PDB domain





Studying Macromolecular Motions in a Database Framework: from Structure to Sequence

1Motions Database

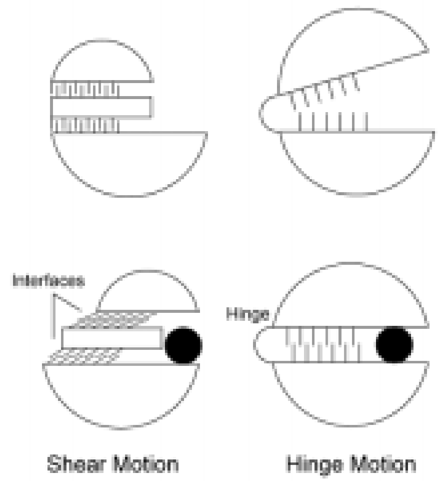
Morph Server, Hinge & Shear, Packing Based Classification

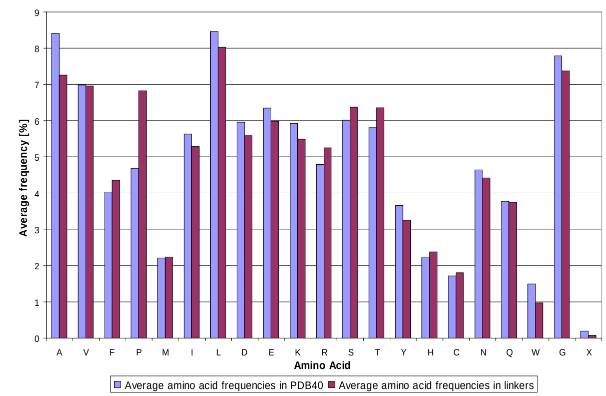
2 Analysis of Packing

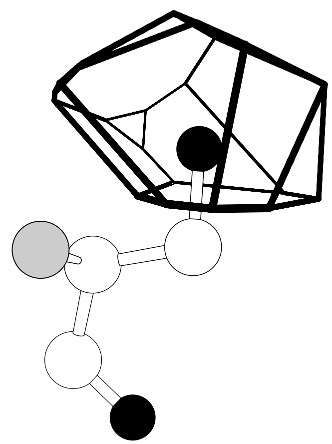
Voronoi Polyhedra, Standard Volumes

3Motion in Sequences

Hinge Profile, Occurrence of Mobility in Genomes







http://bioinfo.mbb.yale.edu/MolMovDB
.../census

Databases: A New and Different Paradigm for Scientific Computing?

Why

- Increasing Amount of Data
  - ◊ PDB at >10K domains and growing
  - ◊ Genome Sequencing
- Papers not sufficient
  - ◊ Difficulty of representing 3D objects on static page
  - ◊ Much important (!) biological information only useful as data to a computer program or a very specialized way to a person reading a paper

What

1 Big calculations on large centralized computers
  - ◊ Aim is prediction based on 1st principles
    — e.g. folding by MD
  - ◊ CPU
  - ◊ Physical Laws

2 Collection of small and interlinked DBs on many different computers
  - ◊ Aim is communication and the discovering of unexpected patterns
    – e.g. HSP ~ hexokinase
  - ◊ Disk + networks
  - ◊ Biological, Statistical, Economic

Studying Macromolecular Motions in a Database Framework: from Structure to Sequence

1Motions Database

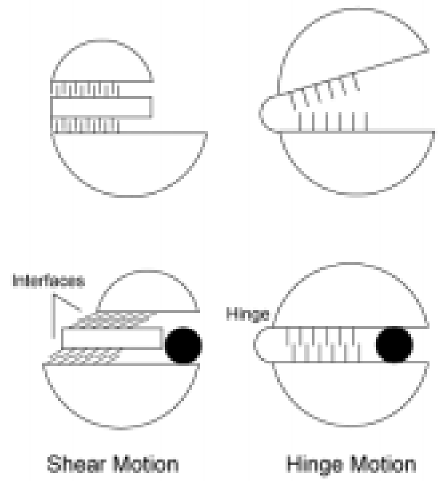
X Morph Server, Hinge & Shear, Packing Based Classification

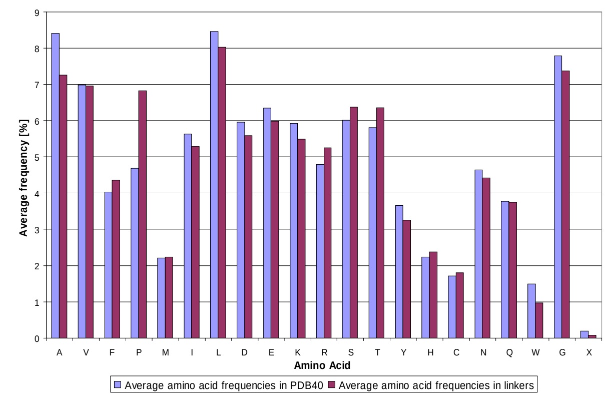
2 Analysis of Packing

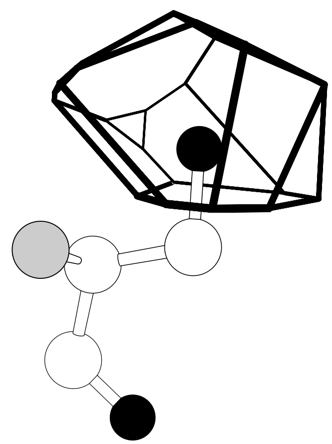
X Voronoi Polyhedra, Standard Volumes

3Motion in Sequences

X Hinge Profile, Occurrence of Mobility in Genomes







W Krebs, R Taylor
R Jansen, J Tsai
T Johnson
C Chothia

http://bioinfo.mbb.yale.edu/MolMovDB

.../census