Digital biology: Relations between data-mining in biological sequences and physical chemistry

L. Ridgway Scott

The Institute for Biophysical Dynamics, the Computation Institute, and the Departments of Computer Science and Mathematics, The University of Chicago, Chicago IL 60637, U.S.A.

This talk is based on joint work with Ariel Ferndandez (Indiana Univ. → Rice Univ.), Steve Berry (U. Chicago), Harold Scheraga (Cornell), and Kristina Rogale Plazonic (Princeton).

1 Overview

Our thesis:
Interaction between physical chemistry and data mining in biophysical data bases is useful.

We give examples to show data mining can lead to new results in physical chemistry significant in biology.

We show that using physical chemistry to look at data provides insights regarding function.

In particular, we review some recent results regarding protein-protein interaction that are based on novel insights about hydrophobic effects. We discuss how these can be used to understand signalling using proteins.

2 A quote

from Nature's Robots ....

The exact and definite determination of life phenomena which are common to plants and animals is only one side of the physiological problem of today. The other side is the construction of a mental picture of the constitution of living matter from these general qualities. In the portion of our work we need the aid of physical chemistry.

Jacques Loeb, The biological problems of today: physiology. Science 7, 154-156 (1897).

so our theme is not so new ....

2.1 Data mining definition

WHATIS.COM: Data mining is sorting through data to identify patterns and establish relationships.

Data mining parameters include:

- Association - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known

Conclusion: Data mining involves looking at data.

2.2 Data mining lens

If data mining is looking at data then

What type of lens do we use?

Alphabetic sequences describe much of biology: DNA, RNA, proteins.

All of these have chemical representations, e.g.,

$$C_{400}H_{620}N_{100}O_{120}P_1S_1$$

All of these have three-dimensional structure.

But structure alone does not explain how they function.

Physical chemistry both simplifies the picture and allows function to be more easily interpreted.

2.3 Sequences can tell a story

Protein sequences

aardvarkateatavisticallyacademicianaccelerative
acetylglycineachievementacidimetricallyacridity
actressadamantadhesivenessadministrativelyadmit
afflictiveafterdinneragrypniaaimlessnessairlift

and DNA sequences

actcatatactagagtacttagacttatactagagcattacttagat

can be studied using automatically determined lexicons.

Joint work with John Goldsmith, Terry Clark, Jing Liu.

2.4 Sequences can tell a story

Protein sequences

aardvarkateatavisticallyacademicianaccelerative
acetylglycineachievementacidimetricallyacridity
actressadamantadhesivenessadministrativelyadmit
afflictiveafterdinneragrypniaaimlessnessairlift

and DNA sequences

actcatatactagagtacttagacttatactagagcattacttagat

can be studied using automatically determined lexicons.

Joint work with John Goldsmith, Terry Clark, Jing Liu.

3 Data mining applied to PChem

Or, what's in all of this for the physical chemist ....

We look at three applications of data mining to physical chemistry:

- microarray hybridization energies are position dependent

  helping to analyze weak genetic signals more accurately

- hydrogen bonds are orientation dependent

  suggesting that molecular dynamics force fields need revising

- peptide bonds are not always planar

  re-writes the rules for protein folding

Data mining provides quantitative predictions for new models.

3.1 cDNA binding

New result:

Energy of binding depends on position as well as neighbor context.

Nature Biotechnology 21, 818–821 (2003)

A model of molecular interactions on short oligonucleotide microarrays

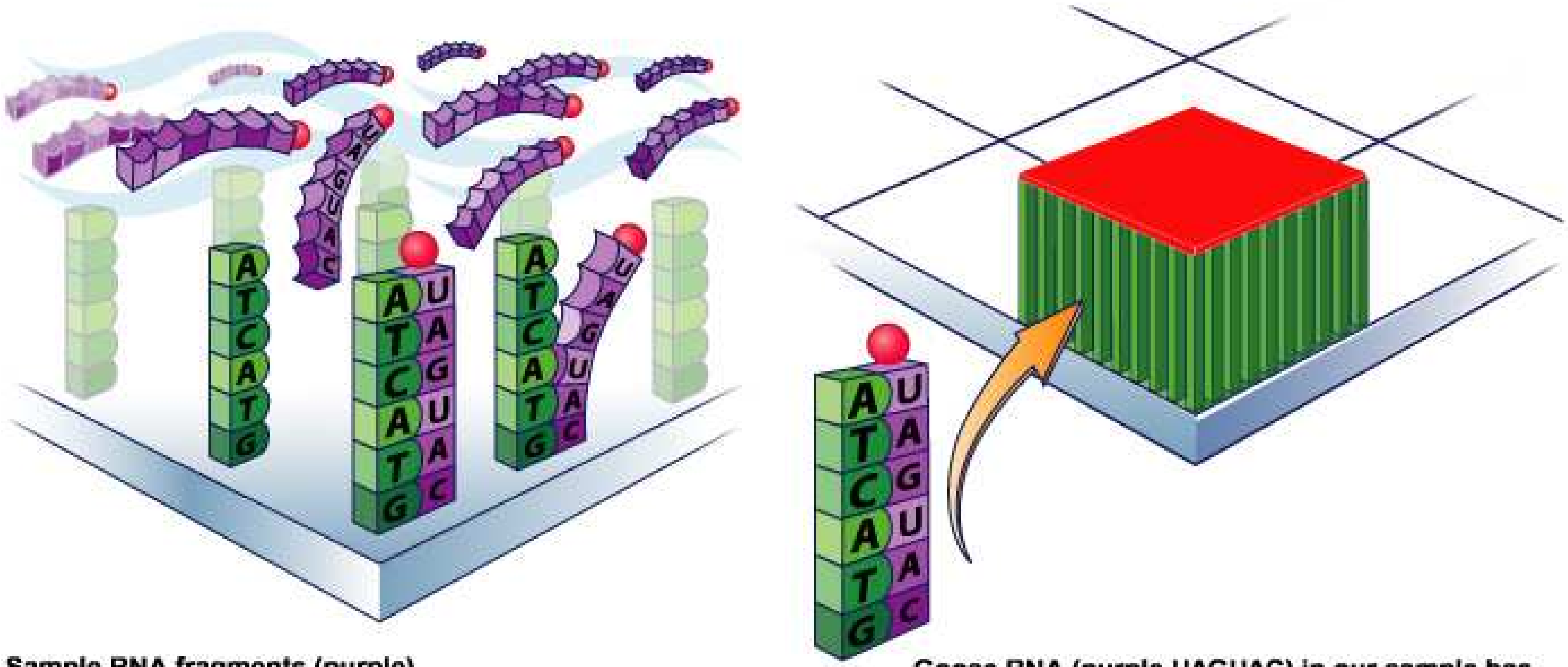
Li Zhang, Michael F Miles & Kenneth D Aldape

PNAS 100, pp. 11237–11242 (2003)

Probe selection for high-density oligonucleotide arrays

Rui Mei, Earl Hubbell, Stefan Bekiranov, Mike Mittmann, Fred C. Christians, Mei-Mei Shen, Gang Lu, Joy Fang, Wei-Min Liu, Tom Ryder, Paul Kaplan, David Kulp, and Teresa A. Webster (Affymetrix, Inc.)

3.1.1 Microarray tutorial (from Affymetrix)

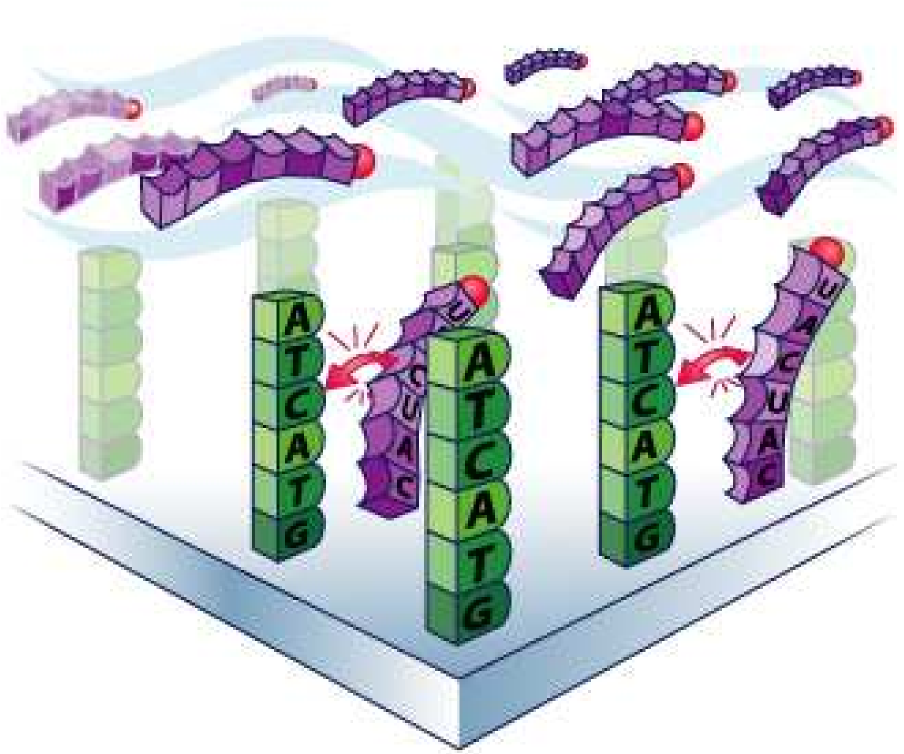


Sample RNA fragments (purple) hybridized to DNA probe array (green)

Goose RNA (purple UAGUAC) in our sample has bound to the goose DNA probe built on the array.

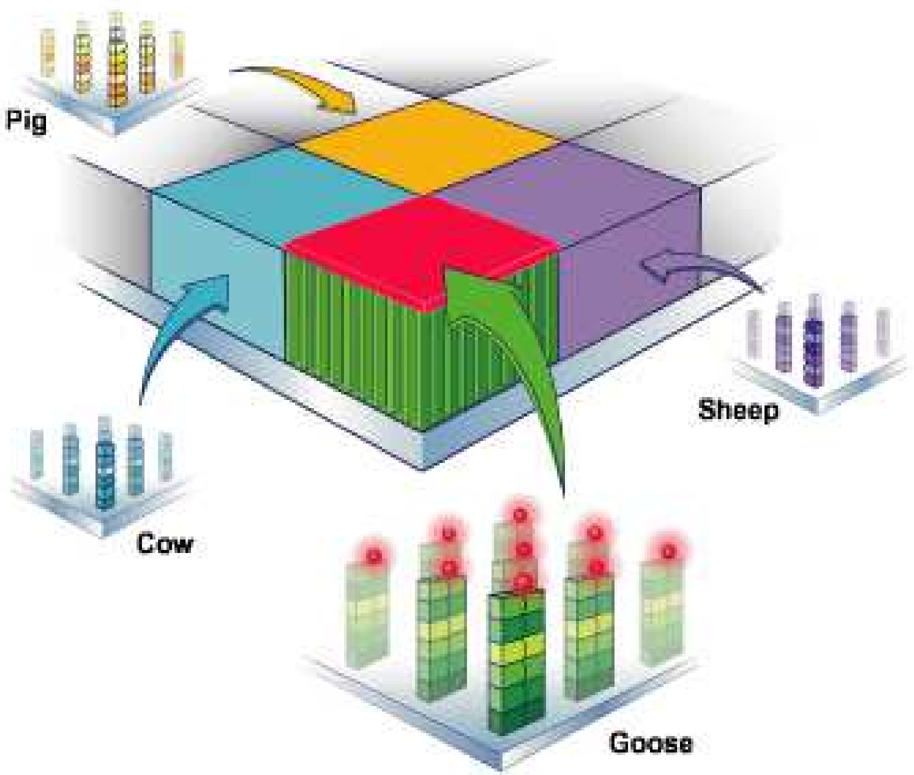
DNA sequences are attached to a slide, and sample RNA is introduced. RNA has flourescent tags added.

3.1.2 Microarray tutorial (from Affymetrix, continued)



C does not stick to another C, so no match is made

Shining a laser light on the FoodExpert ID Array causes the tagged RNA fragments that hybridized to glow



Hmmmm. C does not stick to C; seems reasonable, but maybe we should check. What about G binding to G? A to A? T to T?

3.1.3 Models for RNA/DNA binding strength

For a sequence $\sigma = (\sigma_1, \ldots, \sigma_n)$ (ignore end effects)

Sequence composition model: $\sum_{i=1}^{n} w(\sigma_i)$

Basic nearest-neighbor model: $\sum_{i=2}^{n} W(\sigma_{i-1}, \sigma_i)$

where W is the energy for each pair of letters.

Distance-dependent nearest-neighbor model

$$\sum_{i=2}^{n} d_i W(\sigma_{i-1}, \sigma_i)$$

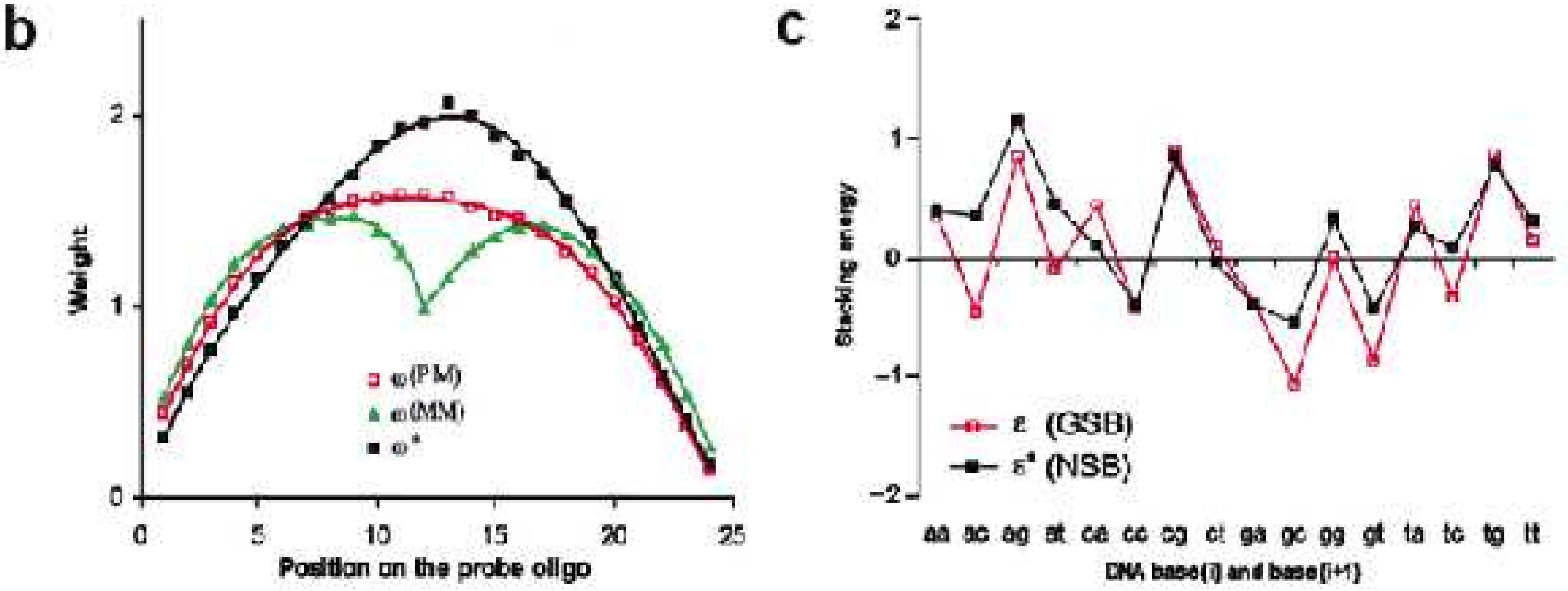
where d_i depends on the position in the sequence.

Another distance-dependent model: $\sum_{i=1}^{n} d_i w(\sigma_i)$

depending only on the sequence composition, not the context.

3.1.4 Using Affymetrix to measure binding

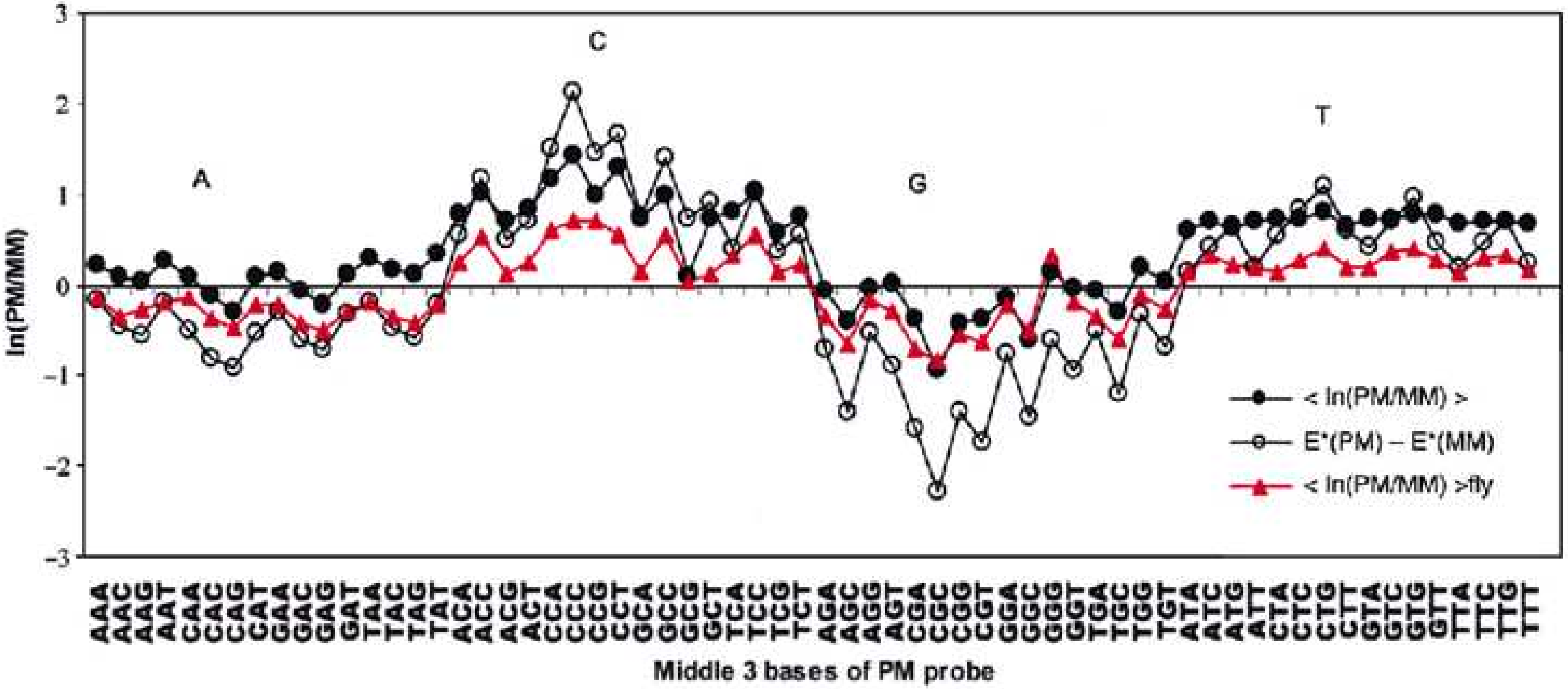
From Nature Biotechnology 21, 818–821 (2003)

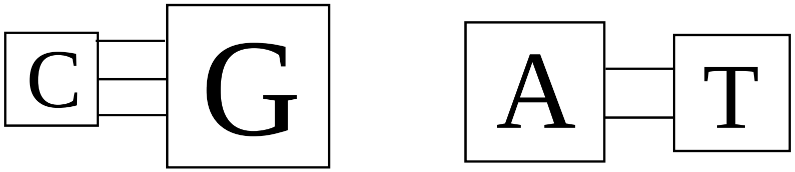


(b) Distance coefficients. (c) Nearest-neighbor stacking energy.

These stacking energies weakly correlated (r = 0.6) with that found in aqueous solution, and are smaller in magnitude.

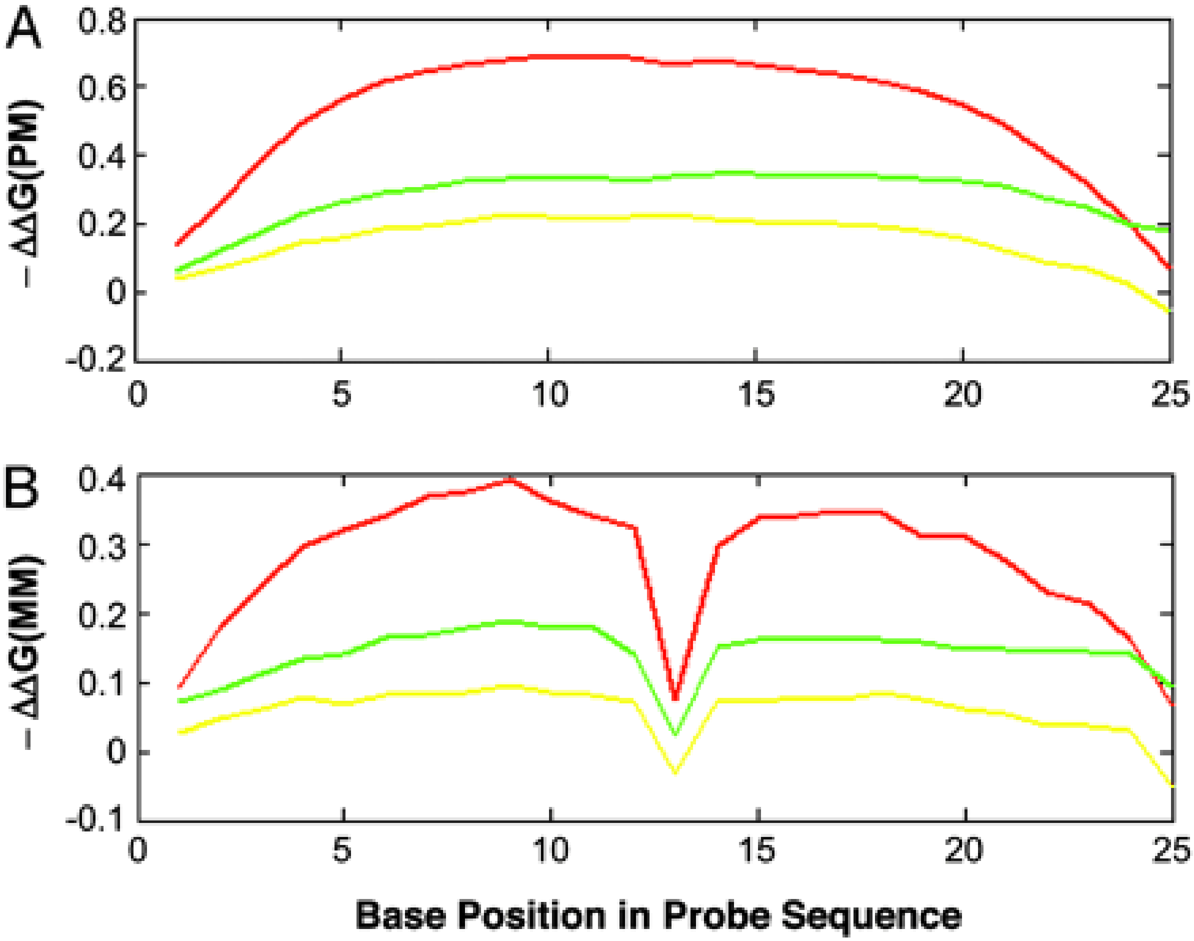
Mismatch signals (C↔G, A↔T) are stronger with certain triplets for non-specific binding (NSB).





DNA pairs differ in **size** and binding strength: removing bulky A or G increases signal.

From PNAS 100, pp. 11237–11242 (2003): model based on bases and locations



The effective $\Delta\Delta G$ values for the 25 probe base positions. The fitted weights ω_{xi} are the effective values for the bases: $x =$ C (red curve), G (green curve), and T (yellow curve) in each sequence position, i ($i = 1, \dots, 25$ from the 3' end of the probe), relative to the reference base, A, in the same position.

Mismatch energies were measured in solution in

Biochemistry. 1999 Mar 23;38(12):3468-77.

Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches.

Peyret N, Seneviratne PA, Allawi HT, SantaLucia J Jr.

Excerpt of abstract: Thermodynamic measurements are reported for 51 DNA duplexes with A.A, C.C, G.G, and T.T single mismatches in all possible Watson-Crick contexts. These measurements were used to test the applicability of the nearest-neighbor model and to calculate the 16 unique nearest-neighbor parameters for the 4 single like with like base mismatches next to a Watson-Crick pair. The observed trend in stabilities of mismatches at 37 degrees C is G.G $>$ T.T $\approx$ A.A $>$ C.C. . . . . The mismatch contribution to duplex stability ranges from -2.22 kcal/mol for GGC.GGC [stabilizing] to +2.66 kcal/mol for ACT.ACT. [destabilizing] ....

3.2 Multiple probes per gene

Affymetrix uses multiple DNA sequence probes

actcatatactagagtacttagact   ctcatatactagagtacttagactt

tcatatactagagtacttagactta   catatactagagtacttagacttat

atatactagagtacttagacttata   tatactagagtacttagacttatac

atactagagtacttagacttatact   tactagagtacttagacttatacta

actagagtacttagacttatactag   ctagagtacttagacttatactaga

tagagtacttagacttatactagag   agagtacttagacttatactagagc

gagtacttagacttatactagagca   agtacttagacttatactagagcat

per gene:

actcatatactagagtacttagacttatactagagcattacttagat

These provide substantial data to assess various binding models.

3.3 Hydrogen bonds are orientation-dependent

Standard force fields in molecular dynamics need improvement.

J Mol Biol 326(4): 1239-59 (2003)

An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes

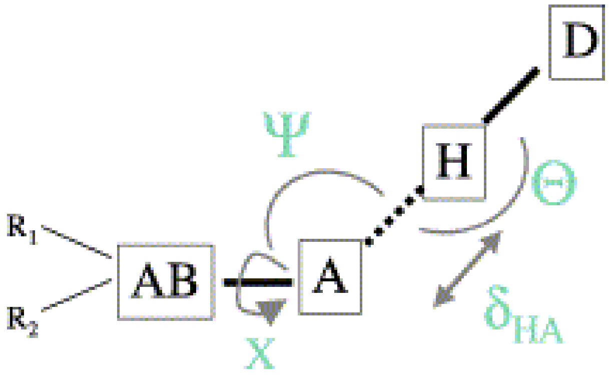
Kortemme, T., A. V. Morozov and D. Baker

and

PNAS 101(18): 6946–6951 (2004)

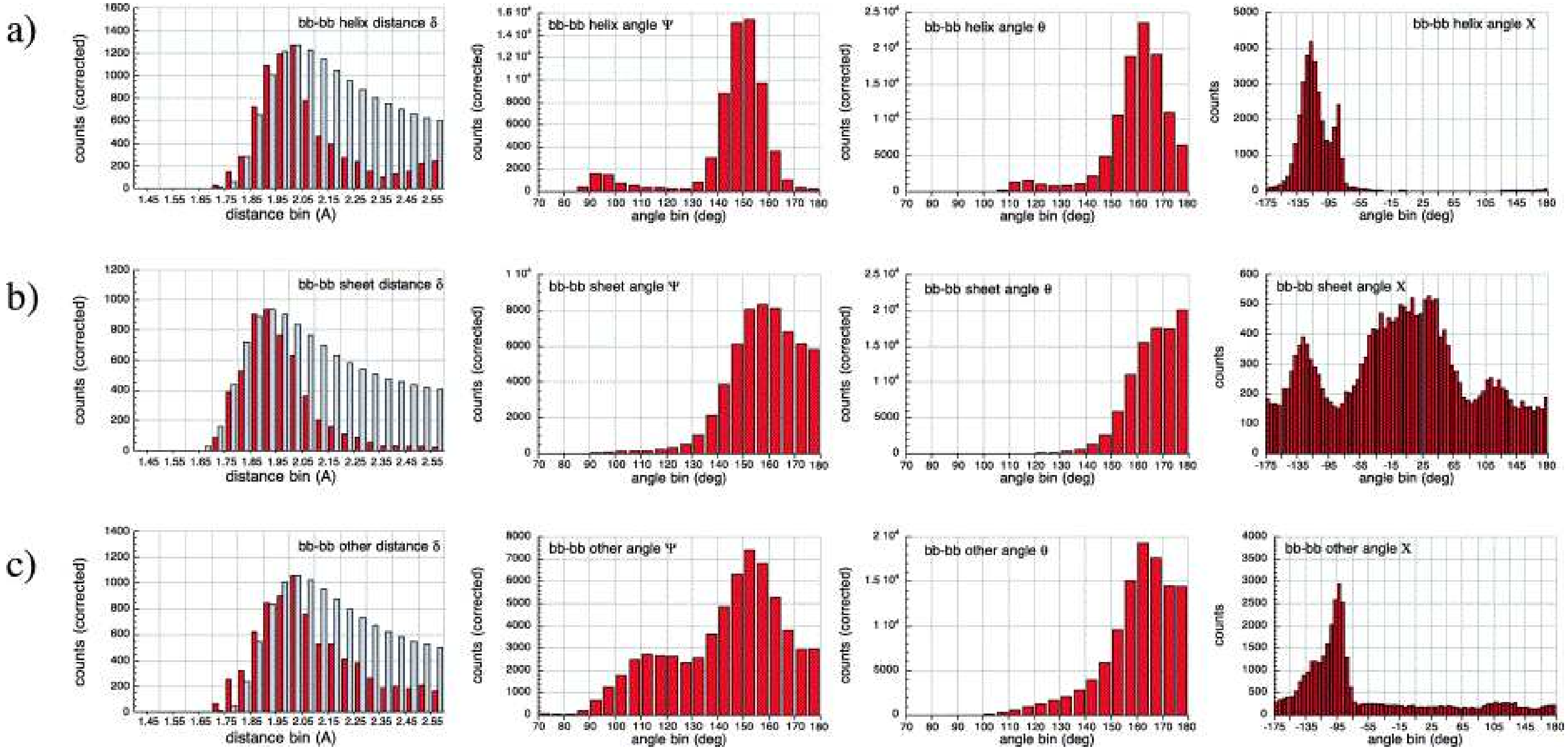
Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations

Alexandre V. Morozov, Tanja Kortemme, Kiril Tsemekhman, and David Baker



Hydrogen bond
distances do not match Lennard-Jones distribution.

Angles are not uniformly distributed.



3.4 Peptide bonds are flexible

Journal of Chemical Physics 121, 11501-11502 (2004)

Buffering the entropic cost of hydrophobic collapse in folding proteins

Ariel Fernández

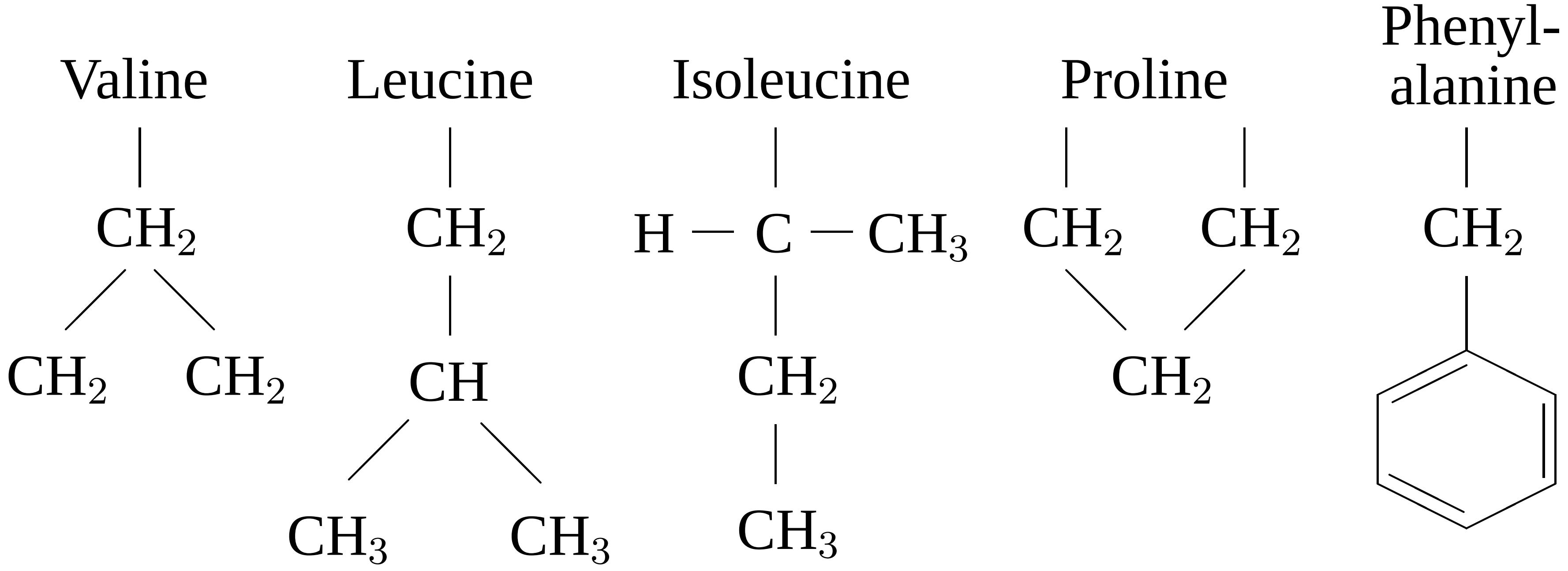
Uses the concept of hydrogen bond wrapping, or dehydration.

- Observes that the electronic environment of peptides determines whether they are rigid or flexible.
- Peptide bond is a resonance between two states: double bonded state depends on polarization.

Peptides can be polarized either by water
or by backbone hydrogen bonds.

3.4.1 Side chains have different properties

Carbonaceous groups on certain side chains are hydrophobic:



Amino acids (side chains only shown) with carbonaceous groups.

3.4.2 Tutorial on hydrophobicity

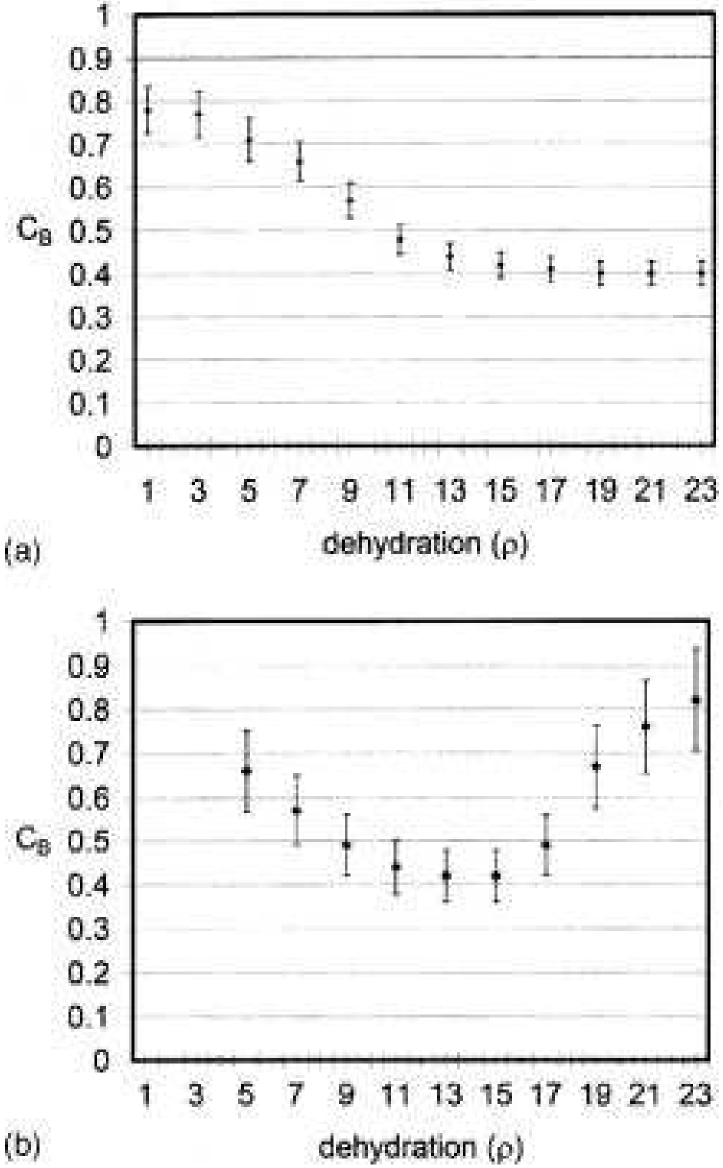
Carbonaceous groups (CH, CH_2, CH_3) are hydrophobic because

- they are non-polar and thus do not attract water strongly
- they are polarizable and thus damp nearby water fluctations

3.4.3 Tutorial on dielectrics

Water removal reduces the dielectric effect and makes electronic bonds stronger.

Number of carbonaceous groups in a region determine extent of water removal and strength of electronic bonds.



From Journal of Chemical Physics 121, 11501-11502 (2004): Fraction of the double-bond (planar) state in the resonance for residues in two different classes

(a) Neither amide nor carbonyl group is engaged in a backbone hydrogen bond. As water is removed, so is polarization of peptide bond.

(b) At least one of the amide or carbonyl groups is engaged in backbone hydrogen bond. As water is removed, hydrogen bond strengthens and increases polarization of peptide bond.

3.4.4 Implications for protein folding

After the “hydrophobic collapse” a protein is compact enough to exclude most water.

- At this stage, few hydrogen bonds have fully formed.
- But most amide and carbonyl groups are protected from water.

The previous figure (a) therefore implies that

Many peptide bonds are flexible in final stage of protein folding.

This effect is not included in current models of protein folding.

Need to allow flexible bonds whose strengths depend on the local electronic environment.

4 PChem applied to data mining

Or, what's in all of this for the bioinformatician ....

We look at three applications of physical chemistry to data mining:

- desolvation helps understand folding rates
- new motif: dehydron=insufficiently desolvated hydrogen bond
- dehydrons are involved in protein interaction
- number of dehydrons correlates with protein interactivity
- number of dehydrons correlates with species complexity

4.1 Determinants of folding rates

Contact order determines folding rates for proteins.

Journal of Molecular Biology 277, 985-994 (1998)

Contact order, transition state placement and the refolding rates of single domain proteins

Kevin W. Plaxcoa, Kim T. Simonsa and David Baker

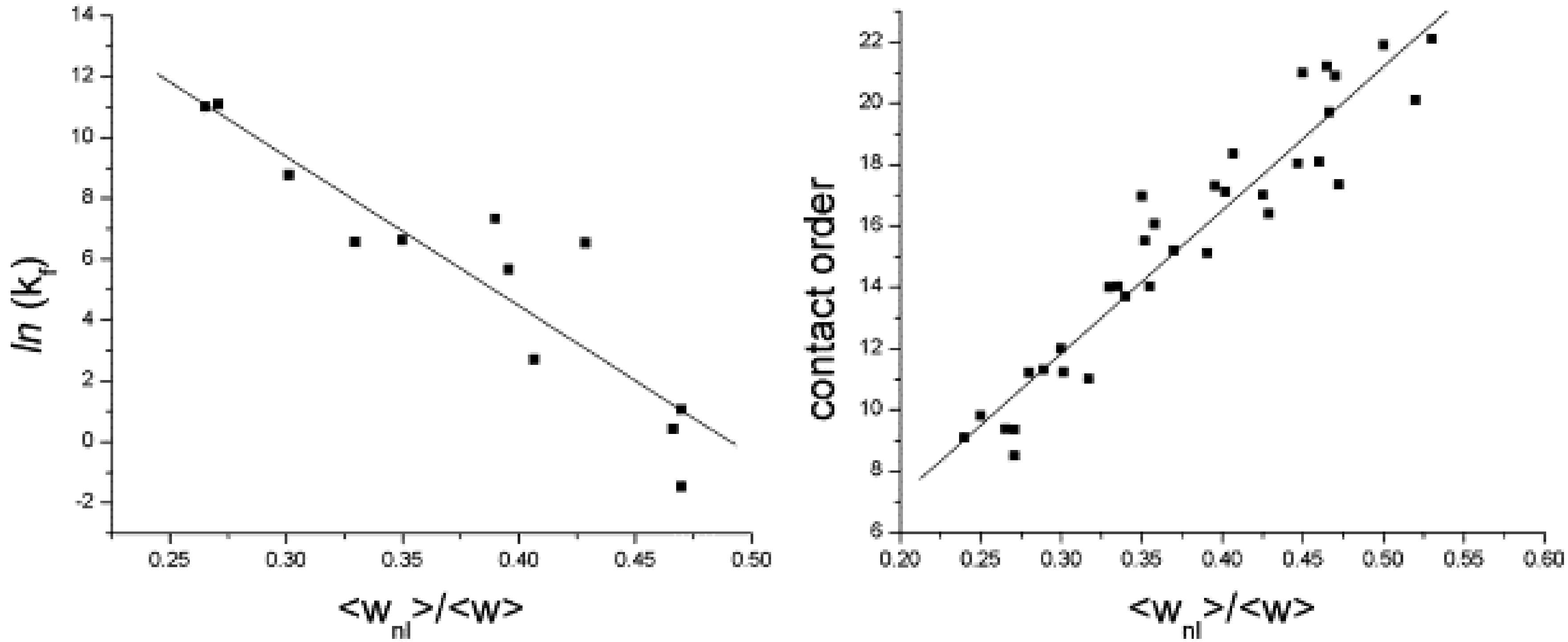
Non-local wrapping of hydrogen bonds gives a similar correlation.

Physics Letters A 321, 263-266 (2004)

Protein folding: a good structure protector is also a good structure seeker

Kristina Rogale and Ariel Fernndez.

From Physics Letters A 321, 263-266 (2004)

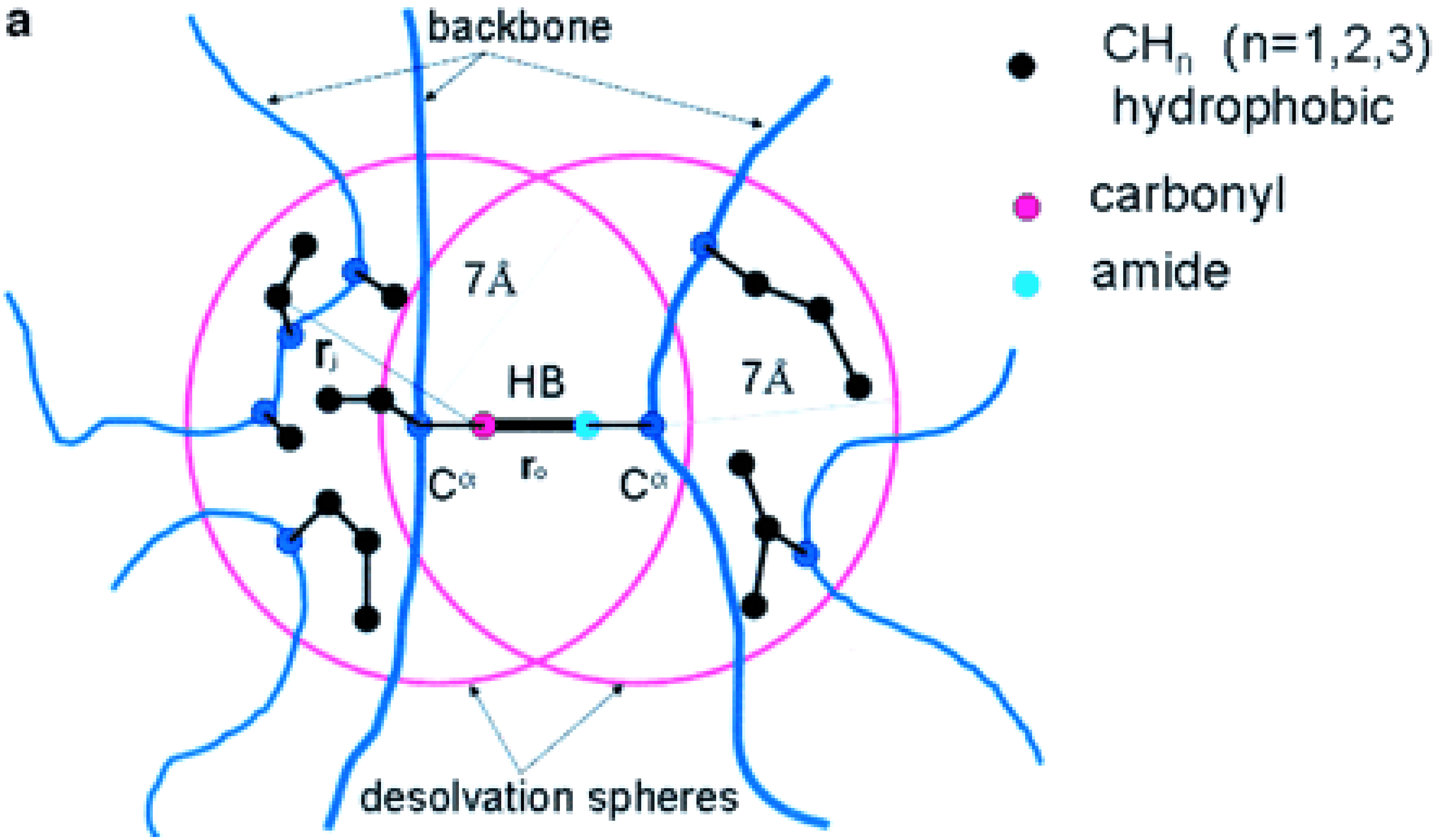


Correlation between the logarithm of the unimolecular folding rate and the average fraction of nonlocal contribution to the wrapping of native hydrogen bonds.

4.2 Understanding wrapping

Hydrogen bonds that are not protected from water may not persist.

Wrapping made quantitative by counting carbonaceous groups in the neighborhood of a hydrogen bond.



4.2.1 Under-wrapped hydrogen bonds

Hydrogen bonds with insufficient wrapping in one context can become well wrapped by a partner.

The hydrogen bond is much stronger when wrapped.

The change in energy makes these hydrogen bonds sticky.

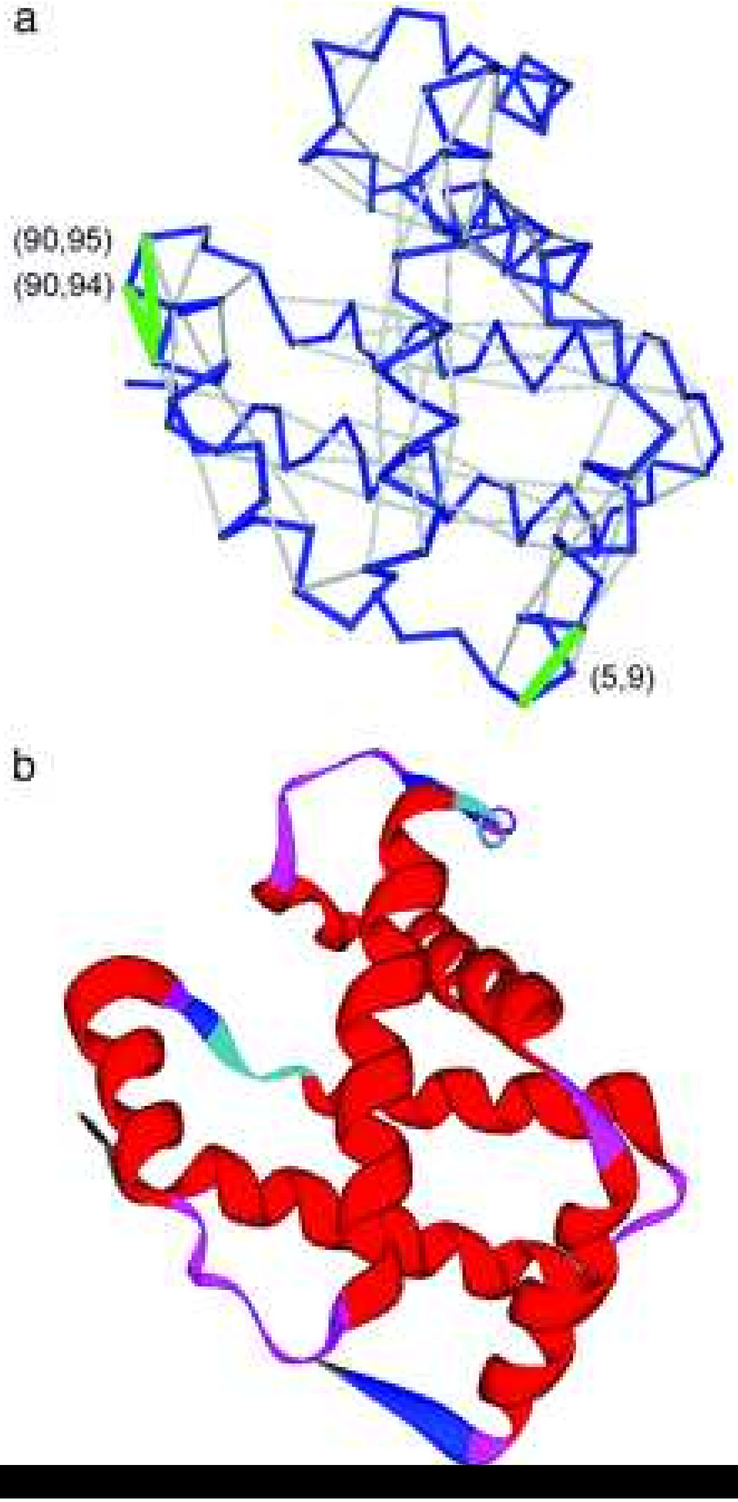
We call such under-wrapped hydrogen bonds

dehydrons

because they can benefit from becoming dehydrated.

The force associated with dehyrdons is not huge, but they can act as a guide in protein-protein association.

In our pictures, we color our dehyrdons green to distinguish from ordinary hydrogen bonds.

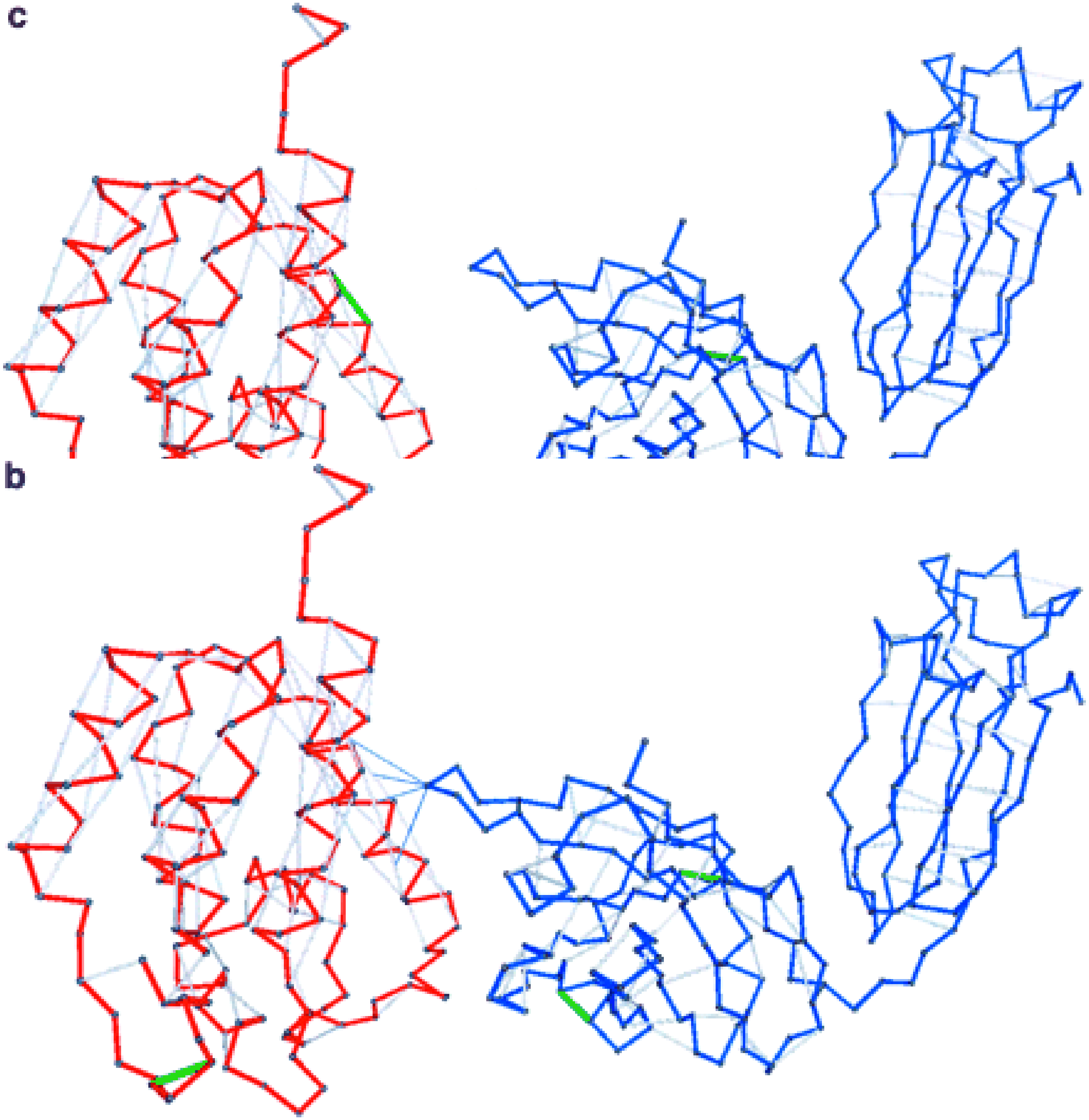


From PNAS

100: 6446-6451 (2003) Ariel Fernandez, Jozsef Kardos, L. Ridgway Scott, Yuji Goto, and R. Stephen Berry. Structural defects and the diagnosis of amyloidogenic propensity.

Well-wrapped hydrogen bonds are grey, and dehydrons are green.

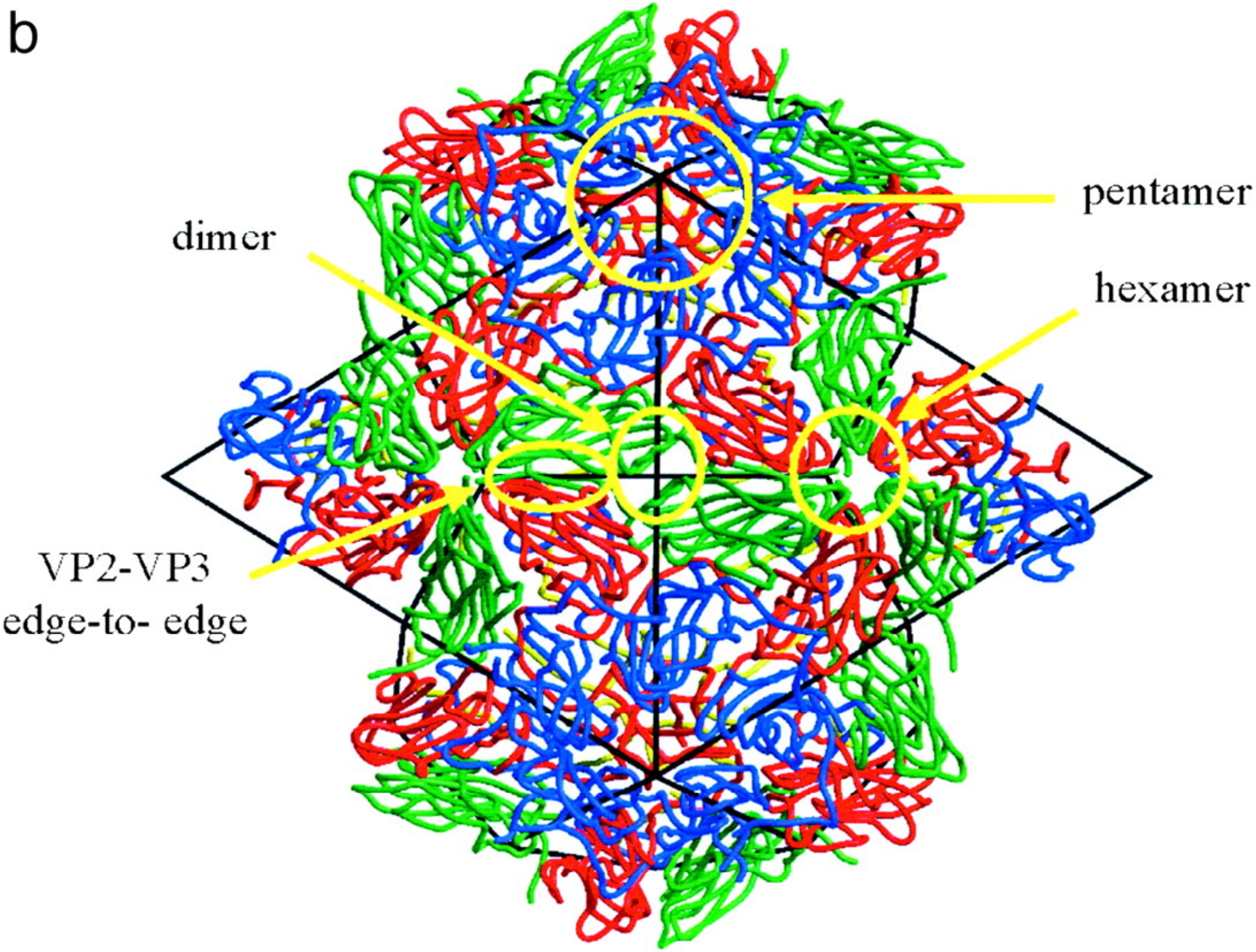
The standard ribbon model of “structure” lacks indicators of electronic propensities.



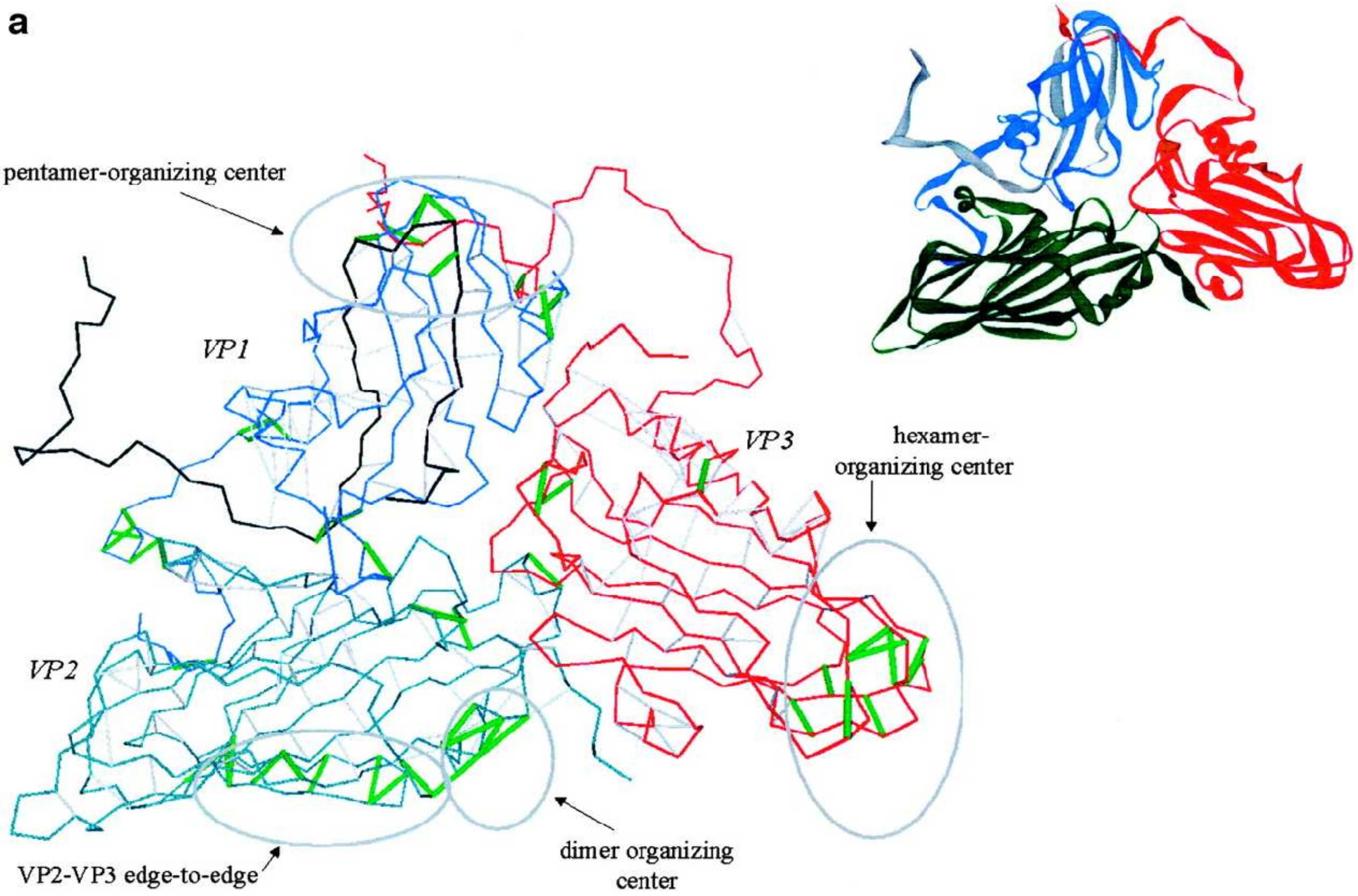
The HIV protease has a dehydron at an antibody binding site.

When the antibody binds at the dehydron, it wraps it with hydrophobic groups.

4.2.2 A model for protein-protein interaction



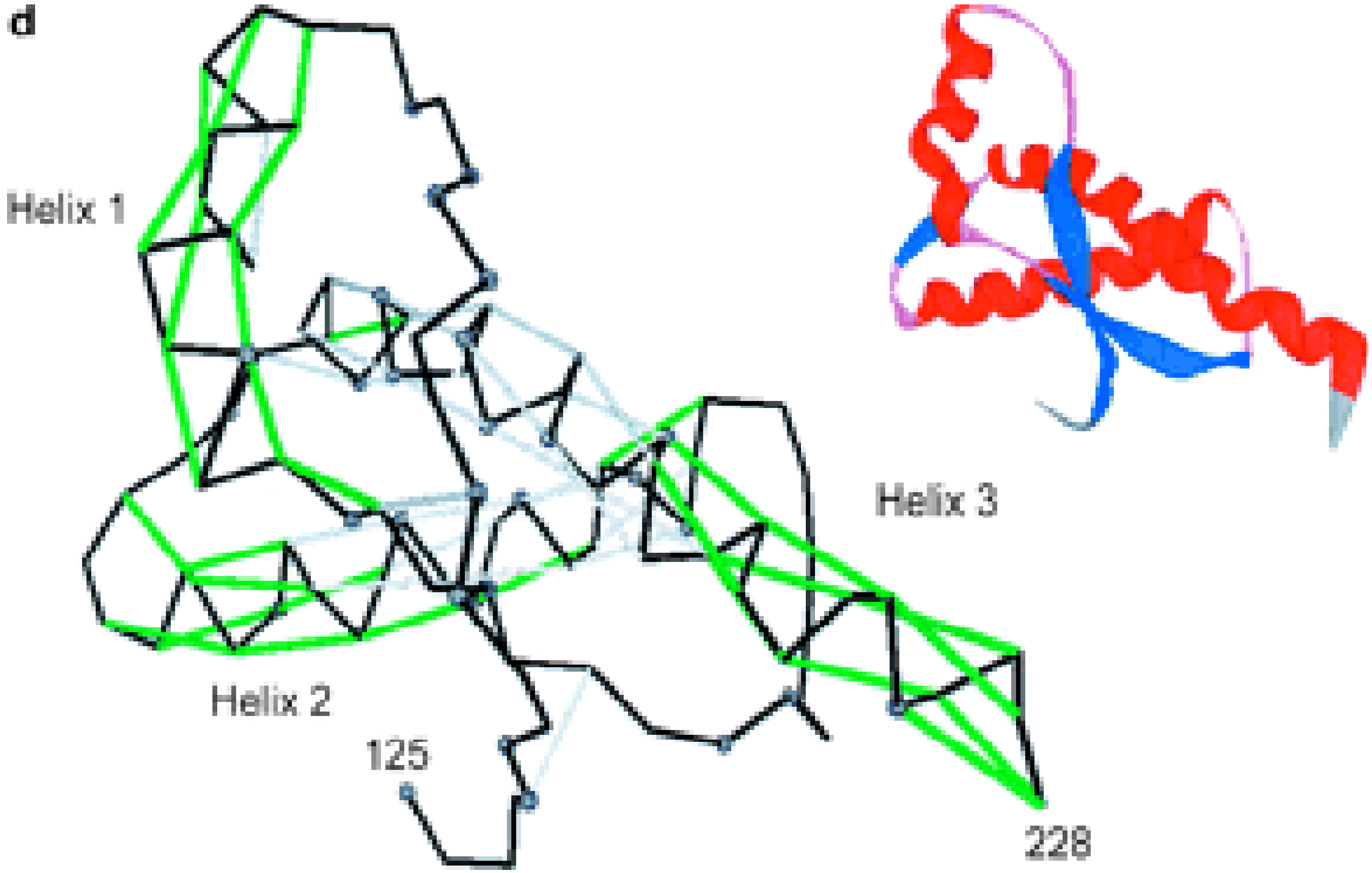
Foot-and-mouth disease virus assembly from small proteins.



Dehydrons guide binding of component proteins VP1, VP2 and VP3 of foot-and-mouth disease virus.

4.2.3 Extreme interaction: amyloid formation

If some is good, more may be better, but too many may be bad.



Too many dehydrons signals trouble: the human prion.

4.3 Dehydrons as indicators of protein interactivity

If dehydrons provide mechanism for proteins to interact, then more interactive proteins should have more dehydrons, and vice versa.

We only expect a correlation since there are (presumably) other ways for proteins to interact.

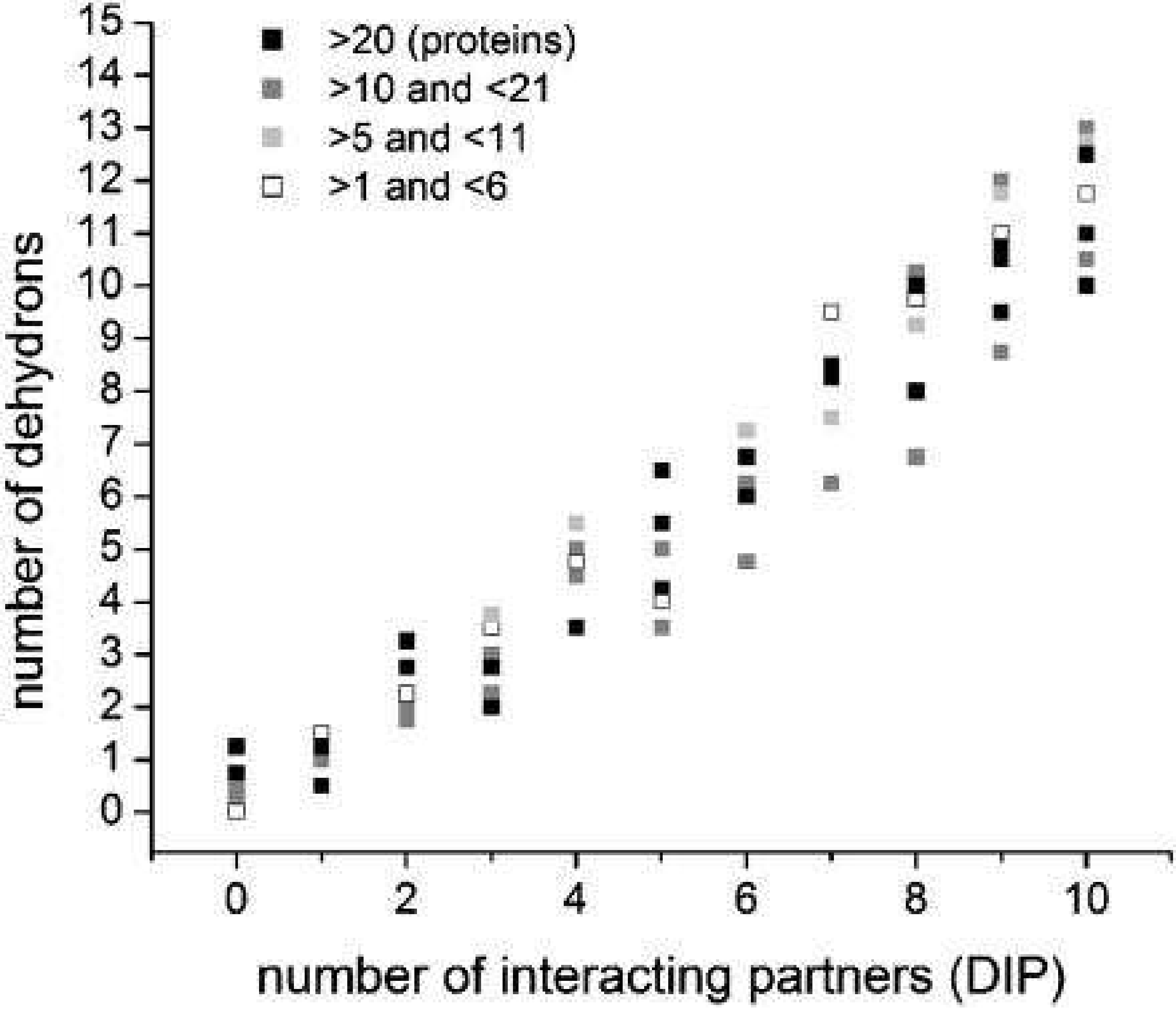
The DIP database collects information about protein interactions, based on individual protein domains: can measure interactivity of different regions of a given protein.

Result: Interactivity of proteins correlates strongly with number of dehydrons.

PNAS 101(9):2823-7 (2004)

The nonconserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks.

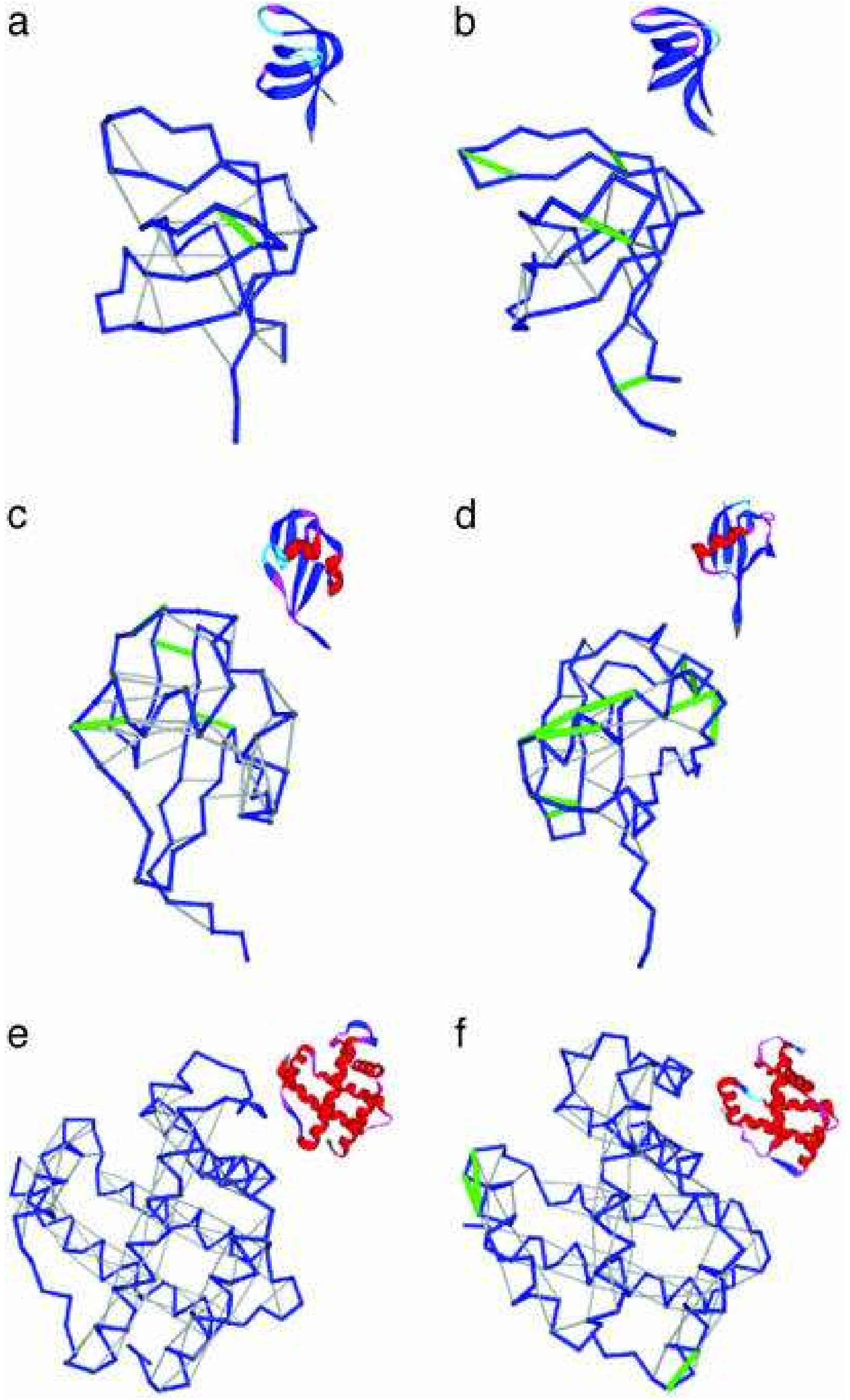
Ariel Fernández, L. R. Scott and R. Steve Berry


■ >20 (proteins)
■ >10 and <21
■ >5 and <11
□ >1 and <6
number of dehydrons
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
number of interacting partners (DIP)
0
2
4
6
8
10

4.3.1 Dehydron variation over different species

Species (common name)	peptides	H bonds	dehydrons
Aplysia limacina (mollusc)	146	106	0
Chironomus thummi thummi (insect)	136	101	3
Thunnus albacares (tuna)	146	110	8
Caretta caretta (sea turtle)	153	110	11
Physeter catodon (whale)	153	113	11
Sus scrofa (pig)	153	113	12
Equus caballus (horse)	152	112	14
Elephas maximus (Asian elephant)	153	115	15
Phoca vitulina (seal)	153	109	16
H. sapiens (human)	146	102	16

Number of dehydrons in Myoglobin of different species



Anecdotal evidence:

the basic
structure is similar, just the
number of dehydrons increases.

SH3 domains are from

nematode C. elegans (a)

H. sapiens (b);

ubiquitin is from
E. coli (c) and H. sapiens (d);

hemoglobin
is from Paramecium
(e). and H. sapiens-subunit (f).

4.3.2 Dehydrons as indicator of interactivity

Is this interactivity an indicator of complexity?

Is this complexity an indicator of evolution?

or is it just Intelligent Design?

The number of dehydrons is greater in more ‘complex’ species.

If this is evolution, then we imagine that protein interactivity became a dominant way to explore biological space, once genome complexity stabilized.

5 Conclusions

The interplay of bio-data mining and physical chemistry can be a productive two-way interaction.

6 Thanks

We are grateful to the Institute for Biophysical Dynamics at the University of Chicago for generous support of this research.

We are also grateful to the developers of the PDB, DIP and other biological data bases.

Thanks to DIMACS for the invitation and logistical support!