May 16, 2012

**Review of prior knowledge**

Liang and Jackups counted pairs in contact (though I don't know exactly what this means - Voronoi cells?) in each of eight areas, and residues in each of eight areas. The areas are: extracellular cap, periplasmic cap, and interior/exterior periplasmic/extracellar side and core.

This allows the presence of a residue or contact to be labeled "typical" or "atypical", with a few different measures:

* Probability that a random object in that region is of that kind
* P(object|region) divided by the probability of mutation into that object (likelihood ration for selection hypothesis vs random evolution hypothesis; best informed after measuing mutation probabilities in that species)
* P(region|object)
* P(object|region) divided by P(object|protein), based on absolute amino acid frequencies. Likelihood ratio for the hypothesis that this region is different vs the hypothesis that all regions are identical, I guess...

From this measure of typicality intuitively follows a measure of energy: assuming that selection pressure is to a large degree for stability of the folded form , greater typicality corresponds with lower energy.

Anfinsen hypothesis: folded form is lower energy. Vik has indicated in my presence that β-amyloid is the lowest energy structure, but that is an aggregate: I suppose Anfinsen's hypothesis is that for an *individual protein* the most stable form is the natural form, and that chaperones catalyze folding and prevent aggregation.

But even from this hypothesis it does not follow that folding-related selection pressure is for stability of the native state.  
Firstly, there's a stability beyond which no more stability is necessary; it really just has to be the *most* stable.   
Secondly, there is negative design: selective pressure to raise the energy of misfolded state.  
Third, foldons. Maity et. al, Englander lab, U Penn studies of Cytochrome C. Cytochrome c has distinct regions that fold and unfold together. So something on the surface of one foldon that catalyzes the folding of a neighboring foldon, for example, will be selected for for making the stable state kinetically accessible.

But maybe that isn't actually the rationale between these knowledge-based potentials. That is, instead of "inference from frequency to energy", there's something more?  
Ezβ makes the assumption that the energies are Boltzmann distributed. Higher frequency, lower energy. That's it. There's just some accounting for entropy introduced by random mutation (which makes me wonder why they bothered putting an entropy constraint on Jim's β-barrel designs; the assumption behind the potential is that entropy is a counterbalance for energy, not that it's functional)

**Russ and Ranganathan**

From Russ and Ranganathan:

"The core problem of rational protein design is the construction of a potential function that describes protein stability. The logical basis for this energy function arises from the thermodynamic hypothesis that the native structure of a protein is given by the conformation of amino acids that minimizes the net free energy of the molecule."

Russ and Rangathanan add another assumption behind knowledge-based potentials: changes in ienergy can be calculated from changes in amino acid identity without changing structure. But on the contrary, a change can cause a structural change, which maybe releases enough free energy to make up for a lost interaction (this could happen if the changed structure would have been high energy with the original sequence *because* of the residues that participate in the originally observed (removed) interaction), or through mechanoical coupling strains a distant weak bond.

So... I guess that just means that... assuming only *small* changes in structure, calculating an energy change requires an energy minimization step.

Support for "balance between stability and entropy" idea: a protein with the consensus sequence has been shown to be hyperstable.

OKAY. So what did I learn from the Russ and Rangathan paper?  
Without an energy minimization step, a statistical potential based upon pairwise contacts is fundamentally innaccurate, due to mechanical coupling and other structural shifts. In effect you are comparing the energy of a true structure of one sequence to a rough homology model of another sequence. (that is, when you are estimating the energy of a mutation.)

There are statistical potentials that focus on stuff besides membrane depth and pairwise contact. Like... binary patterning.

Consensus sequences are hyperstable, lending support to the idea that selection pressure is for stability of native state. (which makes me wonder if these work, but not for the reason people think they do... if the equilibrium is an evolutionary one, not a physical one)

Test of prediction of optimization for low ez-β: the consensus sequence should have the lowest ez-β of them all! Seems a little tautological...  
if all positions are describable just by z, then the consensus will reflect the frequencies, and the frequenceis were used to formulate ez-β. High frequency, low energy, by definition. Cannot then say, low energy, high frequency, that's confirmation! Theoretically it is, but it's just the confirmation that makes up for the complexity of a theory with so many free parameters. You can't count the info used to set your parameters as evidence. (let's say you're fifty percent sure that the price of onions at the grocery store is always the same, and 50% sure that it's not always the same but varies within a range from x0 to x9. Also, an onion is 100% certain to be less than a dollar. You go to the grocery store and see that the onion is $.69. This is stronger evidence for the theory that it's always $.69 than that it's always between .60 and .69. But the theory that it's always $.69 had a lower prior: 100 of these theories with 50% probability mass spread among them, whereas 10 of the range theories with 50% probability mass spread among them. It's ten times more confirmation of a theory with a prior ten times lower)  
yes, because it's the theory plus parameters that gets tested, isn't it?  
But, you can just do a leave one out test! Ignore all the sequences in that cluster when constructing the potential!  
then... isn't this just a statement of uniformity of frequencies at particular depths?  
YES! It is! And Ezβ is the assumption that there exists this uniformity, and its physical cause is lipid interaction energy!  
The question is just... what level of uniformity is significant?

**Naveed, Jackups, Liang 2009 Oligomerization section - prediction from structure**

using TMSIP, prediction of oligomerization and interface is pretty straight forward.

There's some measure of average deviation of strand energy from average strand energy. I think there's a typo and it's really the variance of strand energy. A cutoff of this value is used as prediction of oligomerization state. Clearly ridiculous over the space of all possible inputs (think alternating high and low energy), but seems to work very well: predicts all seven oligomers, and out of the 18 other proteins, returns only one false positive. (sensitivity 100%, selectivity 94%, accuracy 96%)  
Would be more convincing if the *cutoff* was also free in the leave-one-out test, but that's probably a minor mistake.

Prediction of oligomerization *site* is more sensible: find the largest contiguous region of higher-than-average energy strands. Makes no assumption about *size* of the oligomerization site: if the strands have a smooth up and down gradient of energy, the prediction will be that half the protein is an oligomerization site. If the protein is mostly above average energy strands with one *very* low energy strand, the prediction will be that all strands but one are oligomerization sites. The average is a dubious cutoff - they could have done better with a free cutoff like in the prediction of oligomerization state.

Yet, cutoff manipulation probably wouldn't have done much good, since neither the sensitivity nor selectivity are that good (raising cutoff can only increase selectivity, lowering cutoff can only increase sensitivity). P[I|+,O] is 76%, and P[~I|-,O] is 80%. But the prior is reasonably high, so this adds up to a P[I|+,O] of 78%.

**Today I learned**

Naveed's method of predicting oligomerization state works pretty well. It's been used once to predict oligomerization state from sequence, ass part of a characterization of a putative β-barrel membrane protein. The general principle has been noted, and has lent support to another group's thoughts about the determinents of binding. In that case, an energy interpretation of TMSIP ios used, although since interfaces are under different selective pressures possibly for reasons besides maintaining a different structure (different flexibility? How come nobody talks about the *movement* of membrane proteins?) it does not seem at all clear to me that an energy interpretation is appropriate.

Best evidence FOR energy interpretation would be: prediction of energies of course!  
Other evidence: manipulation of the *presence* of oligomerization sites, since I would guess that a stable oligomerization site means stabilizing interactions with *something*. And that certainly is the majority view, so I'm pretty ocnfident in my guess. Though this experiment was done, it was not done well. Not thorough, no elimination of alternate possibilities, and I don't even really see who cares enough to *do* it with that level of thoroughness.  
The paper where they stabilize proteins. That's pretty important evidence. It means that moving it to the highest frequency contact also lowers the energy. The *exceptions* would be interesting to examine: what *else* is there selective pressure for? When a frequency does *not* predict energy.  
**\* there's something very interesting going on here... when you want to know an atypical unknown cause, you look for exceptions to the usual pattern of cause and effect...**

Naveed et. al look at variance of strand energy (if I'm right about that being a typo) to predict oligomerization *state*, and contiguous above-average strand energy to predict oligomerization *interface*. Predict *size* as well as *location* of interface. State predictions much better than interface predictions.

There's a paper on insertions and deletions and how they affect interfaces. Verification that it's stuff that goes on *at* the interface that determines whether it stays an interface.

Consensus sequences are hyperstable - evidence of selective pressure for stability - *strong* evidence that it happens a lot, though not evidence that it happens *always* - since a consensus sequence reveals the trend of evolutionary pressure if there are enough sequences.

Leave-one-out tests are *awesome*. They're basically telling you how predictions from your dataset do on other stuff drawn from the same source, and you still get to use most of your data to make the prediction! You just gotta worry about how they'll perform on data drawn from a *different* source, when the tank pictures weren't all taken on cloudy days.

Alright. So in outline.

* Regarding thermodynamic hypothesis and the thermodynamic interpretation of TMSIP
  + β-amyloid aggregate is most stable form of any protein. According to thermodynamic hypothesis there is still role for chaperones as catalyst and to prevent aggregation
  + Cytochrome c and other proteins have been found to have foldons (units that cooperatively fold and unfold) and foldons help their neighbors fold; suggests that in design,e ven if you *knew* the lowest energy state you may need to design a folding *pathway* to make sure it happens sometime in the next ten thousand years
  + Consensus sequences are hyperstable
  + TMSIP has been used to engineer hyperstable proteins, though I don't exactly trust Liang and Jackups with regards to experimental design considering that their engineering of monomeric OmpF managed to be done in such a way as to be neutral to the energy interpretation of TMSIP
* Regarding Naveed's method
  + olig state predicted with strand energy variance, olig interface predicted as largest contiguous above-average energy segment, olig state predictions sensitive and selective, olig interface predictions only like 80% accurate (but that's better than I could do from sequence for SURE)
* Regarding applications of Liang's work
  + Quoted in introductions, sometimes introduction quoted in introduction
  + Frequency counts informed other work
  + Never seems to actually get *used*, which is unsurprising because they don't make it easy! No software package, no simple equations - maybe one reasonf or the trend to simple theories is jsut that they're easy to learn to use so they spread better! Even if Ezβ has worse predictions at least you can write a program to do the calculations in fifteen minutes
  + Confirms intuitive idea that olig sites are nstable without their binding partner
    - So, if you want to confirm an intuitive idea, see if there are any predictive methods based upon that intuitive idea - in other words, whether the intuitive idea has been formalized and tested! OF COURSE that's beautiful but I never thought about it before! Like... strand length? Any predictive methods based on strand length? Curvature has been *statistically* correlated with *permanence* of an interface - that was a statistical study - so there's statistical studies too, of course...
* Workflow
  + Procrastination: Like half my time since I got started around four o'clock has been work!
    - Before attempting to review what you've just done, sit back and think, about whatever! Preferably work *related­*,but not necessarily directly
    - Make clear goals. Write *questions* before you read something boring... otherwise how will it have value to you? Needs to be answering your questions to have intrinsic value, unless you see some beautiful inferential structure behind it
    - Make clear subgoals for each period of work!
  + Review after working, it feels good for some reason and I get to rehearse what I've learned. Write randomly and then make outlines
    - Final review onlly takes like a half hour