

Wicked Approximations: proteins move like a network of springs

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/19/2008 11:35:31 PM

Proteins are molecules that possess the rare property of self-assembling into compact little machines that move. And not just any kind of jiggling and wiggling motions but clean mechanical motion, like the gears on a bike, or the winch of a crane.

The dynamic nature of proteins was not appreciated for a long time due partly to the difficulty of measure fast protein motions and partly to the prevalence of X-ray structures. For a long time, the principal method for determining the molecular structure of a protein was through the analysis of X-ray diffraction patterns of proteins trapped in crystalline form. This gave the impression that protein molecules existed in discrete static states. In reality, many proteins are highly dynamic structures, where the static structure of an X-ray structure represents an artificially constrained conformation due to the formation of the crystalline lattice.

Today, there are many beautiful NMR experiments that can measure microscopic fluctuations in a protein in solution at room temperature. Work is in progress to measure large motions in proteins directly.

Still, understanding the motion of a protein seems like a perfect fit for computer simulation. There is a pretty solid consensus that molecular dynamics simulations is a reliable method for simulating protein molecules. Sadly the ability to

simulate a protein molecule faithfully using molecular dynamics far outstrips our ability to see any useful motions in our simulation. It just takes too damn long to simulate a protein to the time-scale required to see anything interesting.

Approximations that work well in protein studies are rare but in recent years, a really neat approximation technique has been developed that models really large motions in proteins.

The method is the Gaussian Network Model (GNM) approximation, which has been developed in large part by Ivet Bahar from the University of Pittsburgh.

The pre-cursor to the GNM is Normal Mode Analysis. Normal Mode Analysis is a bastardization of molecular mechanics that uses the classical-mechanics idea that the minimum of any potential can be approximated as a harmonic function. Near the minimum, forces act like springs. One can calculate the springs that approximate the global minimum of a protein using atomic force-fields taken from molecular dynamics simulations.

Unfortunately, the energy minimum of a protein modeled with atomic force fields is really unstable, and produce lots of spurious local minimum.

The genius of the Gaussian Network Model is to strip down the atomic interactions to just the essential ones that allow the protein to be modeled as a clean system of springs. Instead of modeling the full complement of atomic interactions, springs are only

applied between entire residues – either between the C atoms, or between C and C atoms that are in contact.

This level of approximation is sparse enough to produce network of springs that consistently produce clean cooperative motion. The network of springs in a GNM is a tractable mechanical system, amenable to complete analysis. From this network, one can calculate all the modes of vibrations. Each mode consists of the network of springs making one unique vibrational motion. You can classify each mode by its frequency – how long the motion takes to make one bouncy spring motion. It is assumed that the largest, slowest vibration, correspond to the intrinsic global motion of the protein.

There exists a lot of evidence to suggest that these motions are real. One immediate success of the GNM has been to predict the B-factors of many protein structures. In a talk by Valerie Daggett, she remarked that in her unfolding simulations, the unfolding motions invariably resembled motions predicted by GNM's.

Ivet Bahar has built a beautiful web interface to help you calculate the GNM modes of vibration for the protein of your choice, with a java applet animation of course!

The GNM is very good at identifying hinge regions in a large protein. Indeed, there is an even simpler version of the GNM, called FIRST that analyzes only the connectivity of the network of springs to identify the

hinge regions in a protein.

One spectacular example of a GNM calculation is the analysis of the HIV reverse transcriptase [1]. This is a huge, multi-domain complex that moves cooperatively to process RNA to make new DNA. The following is a movie of the motion:

But for me, the most impressive application of GNM is the recent calculation of an entire virus capsid [2]. The virus capsid is made up of hundreds of proteins that form the protective casing for a nasty piece of viral DNA. The virus capsid is known to exist in an immature form with a given symmetry. On maturation, the capsid changes into a different symmetry.

In a tour-de-force calculation, the entire capsid is simulated with GNM, using a clever approximation of simulating only 1 in every 6 residues to generate the network of springs. The slowest mode of this mind-bogglingly massive complex had a collective motion that switched from the symmetry of the immature to the mature form:

References:

- [1]: "Collective dynamics of HIV-1 reverse transcriptase: Examination of Flexibility and Enzyme Function" I. Bahar, B. Erman, R. L. Jernigan, A. R. Atilgan, & D. Covell J. Mol. Biol. 285, 1023-1037, (1999)
- [2]: A.J. Rader, Daniel H. Vlad and Ivet Bahar, Structure (Camb), 2005 Mar;13(3):413-21, 2005.

Hello World: OpenWetWare's Mission

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/20/2008 8:30:39 AM

Hello everyone. I'm Lorrie LeJeune, and two weeks ago I came on board as managing director of OpenWetWare. I've spent most of that time learning about its history, culture, and goals, and I think we're moving in some exciting directions. Over the next few months we'll be working to more fully realize OWW's three-part mission of lowering technical barriers to information sharing in the sciences, building a community of researchers who value open information sharing,

and integrating OWW into the existing publishing and rewards model for science.

One example of how we're working to lower the technical barriers is our new wiki-based lab notebook tool. The OWW Lab Notebook has all the functionality of a paper notebook and more. In particular, your OWW lab notebook is searchable and can display your work both by entry date or by project. We're still shaking out the bugs, but Lab Notebook is currently available to all users. Look for the link in at the top right corner of your window after you log into your OWW account.

We're also exploring both online and

real-world ways of building the OWW community. This includes organizing a series of real-world meet-up sessions, and materials to help communicate the values of open science. If you're interested in attending or hosting a local open science meet-up session, or talking about OWW at a conference, drop a note to me at lorrie@openwetware.org. Finally, we're looking at ways of integrating OpenWetWare into the existing publishing models for science. We want to be able to cite work that's posted on OWW, and eventually we want to explore models for publishing original content. We're

also looking at how to post content like protocols and supplementary materials that often falls between the cracks of traditional academic publishing.

If you have any ideas to help us carry out our mission more effectively, please follow the feedback link in the navigation bar on the OWW main page, or contact me directly at lorrie@openwetware.org. Information sharing in the sciences has come a long way since I worked at the bench and I'm looking forward seeing where the OpenWetWare community can take it.

Another SlideShare plugin from the Wordpress community!

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/20/2008 12:03:07 AM

Many SlideShare users are also WordPress users. In fact, SlideShare

widgets are embedded into Wordpress more than into any other social network. I am constantly amazed by the enthusiasm of the Wordpress community. Joost de Valk has made it even more easy to embed

SlideShare widgets into Wordpress.org. He noticed that we offer a special embed code for Wordpress.com users. His new plugin allows Wordpress.org users to use that same embed code. Just install

this plugin and then you can use the Wordpress.com embed code to embed a SlideShare widget.

Reduccionism and simplification

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/19/2008 5:58:01 AM

This post starts with what might seem as a discussion about computing, but is actually a poor man's discussion about philosophy of science and has nothing to do with computing, it is much more applicable to biology, economics and sociology.

Lets be honest, computer scientists are trained to work for banks and insurance companies, to make web sites, software for cars and things like that. Those domains are actually very simple. A bank might be a gigantic institution, but it is possible to capture, granted with a lot a effort, all its processes inside a computer program. This creates a mental setting: Everything that we need to know is possible to be known: we just decide when to stop.

Now think about those simplistic (this is an understatement) mathematical and computational models for scientific problems (differential equations, Monte Carlo processes, Markov Chains, ...). They model the "important parts" of the issue under study. These models are much simpler than the models working in computers to sustain day to day banking chores. Somehow it strikes me as strange that something as mechanic as a bank needs a more complex model than "nature".

In the context of nature and making mathematical and computational models about it, I have a few things

in mind:

First of all, in many problems in the natural world we don't know what are the important parts to start with. This is very different from the "bank mentality" when you can know everything if you try hard. In my personal case, when I model malarial artesunate resistance, I am modeling something that people speculate how it works, and even if the speculation is correct most of the fundamental parameters are unknown. I am still to read a paper modeling something related to malarial drug use that doesn't have a phrase like: "the relation between this value is and reality is assumed to be this (no citation - or citing something unpublished - or rationale provided)". But the cornerstone of my reasoning is that, in complex processes, the devil is in the details and in the interactions between participating factors (most of which we are unaware of). Soft sciences are holistic by nature. The property of the whole system comes from the everything and everywhere. The "banking" and "hard science" mentality are no good here, we cannot know everything, what we know is probably not enough, and most simplifications will lose something fundamental.

Does this means that I am suggesting that we should stop modeling and all theoretical work? By no means, but we should refocus:

- This is not hard science, don't try to mask it as such. Hard rules,

sensitivity analysis are mostly artifacts to make things look more "serious" and more "demonstrated". This is biology (or even "worse", economy or sociology), you don't c.q.d. here. • Think you can forecast the future? You think you can... then bring me a always correct forecast of the weather in 2 months and I will listen to you. Most models that exist to forecast the future are there because they are very hard to disprove TODAY: climate (as opposed to weather) models, epidemiology, The vast majority of models that can be tested fail (think mathematical finance and the current subprime crisis in the USA, think weather predictions...). • Theoretical work, although not being able predict the future (or explain the past) might help create a cognitive and linguistic framework for discussion: present the fundamental concepts and narratives underlying the research process, make the discourse clearer, less cloudy, point dangerous imprecisions. This is actually the inverse that what happens now: theoreticians speak in a language that most people struggle to understand. • Theoretical work can create interesting questions for field scientists to try to answer: It is the precise inversion of what happens now: We don't want models that are cheated to look realistic. We want reasonable models that fail miserably so that we can ask field scientists: This is failing, why do you think this happens? Have you considered this

other hypothesis? What about testing it?

The existing modeling culture is quite good in the current scientific setting: Makes theoreticians look intelligent with all those complicated mathematics and computer programs (and associated publications) and excuses "practical" scientists of even trying to use their brains: They just apply the existing theory in a process that is more industrial then creative to their research questions. The biggest example that I know of this is phylogenetic analysis: Get data from the field, compute a mutation model from the premise that a small genetic distance is better, burn CPU cycles, publish - You don't even need a human for this - a trained monkey is probably enough.

In economics things are a bit worse: elaborate game theories and such are presented as a "hard, undisputed" justification for an economic theory serving some nice agenda. Nothing more than a authoritarian argument. PS - If you work in an hard science like physics or chemistry you might be thinking that I am smoking something very strong. I don't think that this post applies to hard sciences, that is a different game altogether.

Overcoming data friction

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/20/2008 5:58:10 AM

This headline from Adrian Holovaty's blog speaks volumes about the state of online data in 2008: EveryBlock hiring a Python screen-scraping expert. The recently-launched EveryBlock, a generalization of ChicagoCrime.org, extends that model to other cities and to a broader range of data types. I interviewed Adrian this week for an upcoming ITConversations show, and he confirmed that while some structured data sources are available from the first three EveryBlock cities — Chicago, San Francisco, and New York — the bulk of the data comes from scraping web pages.

One day soon, the person who lands that job will find himself or herself having this conversation at a cocktail party:

Friend: So, what do you do in this new job?

Screen Scraper: I write software to extract data from websites.

F: Where does the data come from?
S: It's in a database. The website's software reads the database and turns it into web pages.

F: So somebody got paid to write software to turn the database into web pages, and now you're getting paid to write software that turns those web pages back into a database?

S: Yeah, basically.

F: So if they just gave you the database you'd be out of a job?

S: No. I'd have a much more interesting job. I'd be able to spend more time finding useful patterns in the data, and writing software to enable other people to find useful patterns in the data.

The irony is that I'd be great at that job. For me, web screen-scraping provides the kind of challenge that other people get from, say, solving crossword puzzles. But it's not the highest and best use of anyone's time. Data friction can be intentional or not. When it's intentional, you might have to file a FOIA request to get it. But in a lot of cases, it's unintentional. The data is public, and intended to be widely seen and used, but isn't readily reusable.

Consider the following two restaurant inspection records for Bully's Deli in New York:

1. in the NYC Department of Health website
2. in EveryBlock

It's the same data, from the same source, but EveryBlock makes better use of it. In the NYC website, you can search by ZIP code and number of violations. In EveryBlock you can search more powerfully, and you can ask and answer questions that matter to you. Maybe you care about shellfish. Have any Manhattan restaurants been cited recently for use

of unapproved shellfish? Yes: five since January 21.

What EveryBlock is doing is completely aligned with the interests of the NYC Department of Health. Tax dollars are paying for those restaurant inspections. The information is published in order to make New York a safer and healthier place. It's great to have this data available in any form, and it's great to see EveryBlock adding value to it. Now it's time to grease the wheels.

Here's one way that can happen. An enlightened city government can decide to publish this kind of data in a reusable way. I've written extensively about Washington DC's groundbreaking DCStat program which does exactly that. I can't wait to see what happens when EveryBlock goes to Washington. But city governments shouldn't have to go out of their way to provide web-facing data services and feeds. Databases should natively support them. That's the idea behind Astoria (ADO.NET Services), which is discussed in this interview with Pablo Castro. If the NYC Department of Health had that kind of access layer sitting on top of its database, it wouldn't put EveryBlock's screen-scraeper out of a job, it would just make that job a whole lot more interesting and effective.

Most Tasmanians unhappy at work - ABC Online

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/19/2008 7:48:45 PM

Most Tasmanians unhappy at work ABC Online - 6 hours ago

A national employment survey has found that the majority of Tasmanians are dissatisfied with their jobs. Online employment agency CareerOne conducted the survey, researching earnings, rewards and promotions.

Tasmanians unhappiest workersThe Mercury We work in an unhappy stateCourier Mail all 4 news articles

Duty Calls

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/19/2008 9:00:00 PM

Yahoo: Living up to the talk on Hadoop

By neilfws (neilfws' shared items in Google Reader)

Submitted at 2/20/2008 1:54:37 PM

I admit that I am a "Google guy". i.e. I like the companies products and their quirkyness. I have also long abandoned a number of core Yahoo products (IM, mail, MyYahoo, etc) due to their design choices and a general preference for most things Goog. However, Yahoo holds a soft spot in my heart, partly due to their history, and the fact that some close friends of mine worked there till recently, but mostly for their open source/*Nix ethos, their ownership of flickr and del.icio.us (two of THE best web properties out there), and for their support of projects like Hadoop

(and for Jeremy Zawodny, who writes one of the best blogs in the business). Yesterday, amidst all the Microsoft morass, Yahoo gave us a picture of what they have done with Hadoop, in some sense a necessity given what Google is doing with MapReduce, but an excellent example of a company making good technology choices. Some of the stats for their new webmap are astounding.

I wonder what it would take to create a map of the scientific web, or subsets of the scientific web and will anyone ever do it.

Yahoo is also hosting the Hadoop Summit in March
Technorati Tags: Yahoo, Webmap, Hadoop