THE PANDA'S THUMB

Update on PNAS flagellum paper

Nick Matzke posted Entry 3081 on April 20, 2007 02:28 AM.

Trackback URL: http://www.pandasthumb.org/cgi-bin/mt/mt-tb.fcgi/3071

Those who have been following the comments section of the first post on the PNAS flagellum paper, entitled "Stepwise formation of the bacterial flagellar system," will see that there have been several developments: ScienceNOW at *Science* magazine has uncritically reported the PNAS paper's all-flagellar-genes-came-from-one conclusion; Behe and other IDers are getting into the act, although they are so clueless they don't really even understand why the PNAS paper is problematic; and PZ Myers and I have dropped hints that several of us PT bloggers are reaching the conclusion that this paper is looking worse, not better, after close examination. We will have more on the technical methodology issues in the next few days. For the moment I would just like to offer a simple response to some comments, and a simple but powerful reason that the "all core flagellum genes are descended from one ancestral gene" does not work.

First, the comments. Some commentators have reacted along the following lines: (1) maybe the paper isn't so bad, just speculative; and/or (2) maybe I've misread the paper or its conclusion was poorly worded, and maybe the authors just meant to argue that *some* of the 24 core flagellar proteins were related, not all 24 proteins.

Unfortunately – and I mean unfortunately because I wish one of these options was true – neither idea is a supportable interpretation of the authors' views. Have another look at Figure 3 from the Liu & Ochman paper:

Navigation

Main
Weekly Archive
Category: Flagellum evolution
« Best Protest Signs. Ever.
» You have the brains of a worm...
Comments (63)
Post a Comment
TrackBacks (1)
Email
Printer-friendly version

Disclaimer

Authors are solely responsible for the content of their articles on PandasThumb.org. Linked material is the responsibility of the party who created it. Commenters are responsible for the content of comments. The opinions expressed in articles, linked materials, and comments are not necessarily those of PandasThumb.org. See our full disclaimer.

Recent Comments

Nick (Matzke) on April 24, 2007 2:40 AM
Dan Gaston on April 23, 2007 4:42 PM
Dan Gaston on April 23, 2007 4:37 PM
Douglas Theobald on April 23, 2007 3:32 PM
Douglas Theobald on April 23, 2007 3:12 PM
Dan Gaston on April 23, 2007 2:47 PM
Reed A. Cartwright on April 23, 2007 1:13 PM
Douglas Theobald on April 23, 2007 12:53 PM
Dan Gaston on April 23, 2007 7:25 AM
Nick (Matzke) on April 23, 2007 3:45 AM

Recent Trackbacks

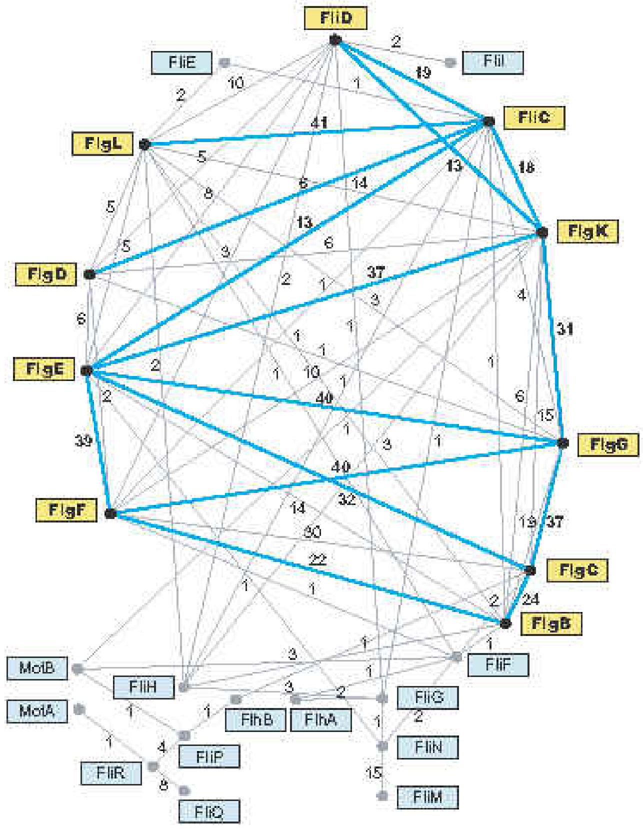


Fig. 3. Network of relationships among flagellar core proteins. Above each link is the number of genomes for which homology between a particular protein pair was detected by pairwise comparison at a cutoff value of 10^{-4} or lower. Blue lines linking yellow-boxed proteins portray the homology network revealed when core proteins of *E. coli* were subjected to pairwise comparisons.

- The boxes contain the names of all 24 core flagellar proteins that the authors identified.
- Each line represents a significant homology hit using pairwise BLAST. The claimed e-value cutoff for significance is 10^{-4} (e=0.0001), which is pretty conservative – researchers sometimes assign homology pretty confidently on e-values up to 10^{-2} (e=0.01).
- The numbers represent the number of times, in 41 tested bacterial genomes, that the homology was found. For example, the already well-known homology between FliC (flagellin) and FlgL (flagellin/hook adapter) was recovered 41/41 times.
- The blue lines represent homologies retrieved in *E. coli* K12, the grey lines hits retrieve in any of the 41 genomes.

In some ways, Figure 3 confirms conventional wisdom. The axial filament of the flagellum – the tube that extends from the basal body out to the filament – is made up of a series of proteins, approximately in this arrangement, with these copy numbers:

~9 copies of FliE (FliF-rod linker)
~6 copies of FlgB (inner rod)
~6 copies of FlgC (middle rod)
~6 copies of FlgF (middle rod)
~26 copies of FlgG (outer rod)
~130 copies of FlgE (hook – the curvy part in most flagellum diagrams)
~11 copies of FlgK (first hook-filament linker)

The Panda's Thumb: Flagellum evolution kerfluffle continued

Site Traffic

SITE METER
4,978,567

Recommend this entry to a friend

Friend's email address:

Your email address:

Message (optional):

Send

> ~11 copies of FlgL (second hook-filament linker)
> ~20,000 copies of FliC (flagellin, makes up the flagellar filament)

The long tube that these proteins form is called the "axial filament" and the proteins as a group are "axial proteins."

Now, if you remember Table 1 from Pallen and Matzke 2006, a simple protein BLAST search will return two groups of relatives within the axial proteins:

- FlgBCFGEK, the rod+hook+first linker protein, is one related group
- FlgL (the 2nd linker) + FliC (flagellin) is the second related group

Figure 3 supports these already-known relationships, which are all supported by 24 or more hits, usually 30 or more hits.

However, the point of Liu & Ochman's Figure 3is that not just the axial proteins, but *all* 24 proteins are connected by one or more homologies to other proteins. Just to make sure the authors really mean this, let's examine relevant bits of the paper.

It is kind of strange to read through the Liu and Ochman paper – you can almost watch their thesis mutate from an interesting, if unoriginal, observation about axial proteins, to a sweeping, wildly unsupportable conclusion about all core flagellar proteins:

> [Abstract] These results show that core components of the bacterial flagellum originated through the successive duplication and modification of a few, or perhaps even a single, precursor gene.

OK, here, maybe they just mean *some* of the core flagellar components came from a single common ancestor.

> [End of the introduction] Our results show that flagellum [should be "the flagellum", sic] originated very early, before the diversification of contemporary bacterial phyla, and evolved in a stepwise fashion through a series of gene duplication, loss and transfer events.

This is not wildly wrong as stated, although the phylogenetic conclusion about flagella being in the last common ancestor of bacteria is disputed, e.g. by Cavalier-Smith. Moving on:

> **Core Flagellar Proteins Arose Through the Duplication and Diversification of a Single Precursor.**

That is pretty clear right there in the section heading: "Single Precursor."

> **Core Flagellar Proteins Arose Through the Duplication and Diversification of a Single Precursor.** When each of the 24 core flagellar proteins of *E. coli* are compared (via BLAST) to all proteins encoded in the *E. coli* genome, their best and often only hits are to other core flagellar proteins. This pattern indicates that the structural genes specifying the portion of flagellum residing outside of cytoplasmic membrane (i.e., the rod, hook, and filament) are paralogs and were derived from one another through duplications.

Here they are just talking about the *E. coli* K12 genome. We have already discussed the axial proteins, so you know this part is confirming conventional wisdom (connecting the FliC+FlgL group to the FlgBCFGEK group is a controversial addition, but that problem is small potatoes at the moment).

Let’s look at Figure 3 again.

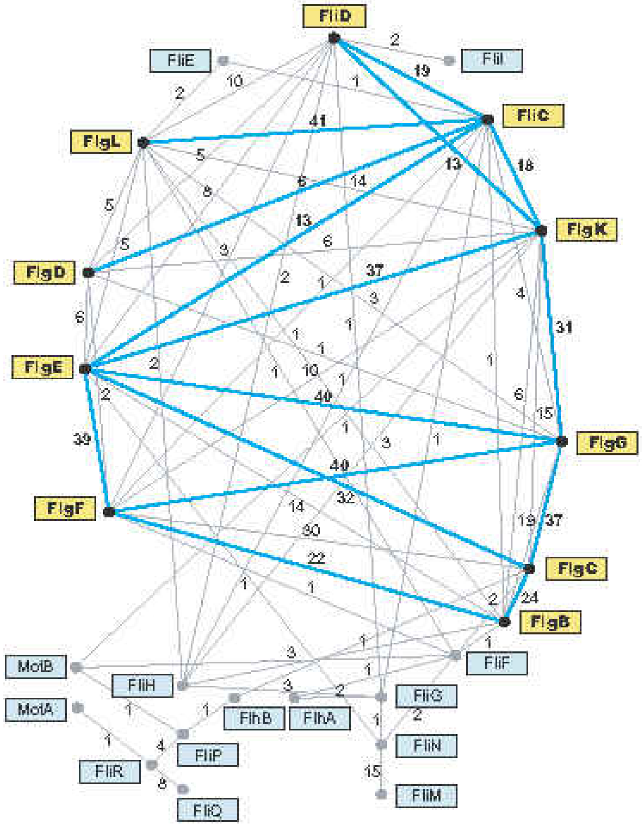


Fig. 3. Network of relationships among flagellar core proteins. Above each link is the number of genomes for which homology between a particular protein pair was detected by pairwise comparison at a cutoff value of 10^{-4} or lower. Blue lines linking yellow-boxed proteins portray the homology network revealed when core proteins of *E. coli* were subjected to pairwise comparisons.

The *E. coli* hits are the blue lines, and the matched E. coli proteins are the yellow boxes. I count ~~14~~ 15 blue lines linking 10 yellow boxes, all axial proteins.

The paper continues on, to look for homologies outside of the flagellum, inside the *E. coli* K12 genome:

> Aside from these matches to other core proteins, pairwise comparisons of these flagellar proteins to the >4,000 nonflagellar proteins encoded by the entire *E. coli* genome recovered cumulatively a total of only 24 hits that reached the same level of significance. [...]

We will skip over some details which are vague and problematic for other reasons, but which are, again, small potatoes. The conclusion of their search of *E. coli*:

> Thus, we conclude that despite their antiquity, the similarities among core proteins to one another are more common and, on average, stronger than to nonflagellar proteins.

It is none too clear, but I think what they are trying to say here is that with the E. coli K-12 genome, they got:

(a) ~~14~~ 15 homologies found between 10 flagellar proteins (out of 24 total flagellar proteins)

(b) and 24 homologies found between ?? flagellar proteins (they don't say how many) and 4,000 nonflagellar proteins

Because (a) is a higher proportion than (b), the authors say that "on average" the within-flagellum homologies are more common. OK I guess. This doesn't establish anything about *all* the flagellar proteins evolving via internal duplications – if you've got external homologies then you have evidence that flagellar proteins could have once had an nonflagellar ancestors.

What about the other 40 genomes? The authors continue:

> We repeated this analysis and compared the core gene set of each other flagellated bacterium to all proteins encoded in the corresponding genomes and among themselves, and we obtained a similar result, i.e., the best (and often the only) hits of the flagellar core genes were to other flagellar core genes.

This is just more of the axial protein homology detection.

> However, by extending this analysis beyond *E. coli*, the similarity-relationships and links among several other core genes were resolved. For example, a highly significant match between fliM and fliN (that was not detected for E. coli homologs) was evident in 15 genomes from diverse bacterial subdivisions (Fig. 3).

The FliM-FliN homology has been well-known for quite a while. Apparently they don't catch the well-known homology between the chemotaxis protein CheC and FliM because they don't do a general BLAST search of a full database. (FliM has two main domains, homologous to CheC and FliN respectively.)

> Cumulatively, each of the 24 core genes shows significant similarly to one or more of the other core genes (Fig. 3), a pattern that would result from their successive origination from one another by independent gene duplications and/or gene fusions.

See, I wasn't making it up. They really are basing their all-from-one-gene conclusion on Figure 3.

> Comparisons of the complete genome sequences of flagellated bacteria revealed that the flagellum is based on an ancestral set of 24 core genes for which homologs are present in genomes of all bacterial phyla. The most striking finding from our analysis is that these core genes originated from one another through a series of duplications, an inference based on the fact that they still retain significant sequence homology.

"[T]hese core genes originated from one another" – pretty clear what they meant.

> Although sequence similarities among some of the rod and hook proteins were noted in early analyses (24), the degree of paralogy for the ancestral set of flagellar genes, and its implications for the origins of the bacterial flagellum, have gone unrecognized.

Um, right. That's because the paralogy has *not* been detected by anyone else, except for the axial proteins and FliN-FliM. Figure 3 reports about 50 new homology matches (the grey lines plus a few blue lines) for ~~14~~ 15 flagellar proteins. Authors and reviewers might have pondered profitably the question of why no one else was able to BLAST up this cornucopia of results before.

> [Summary]
>
> To ascertain the ancestry of the flagellar core genes, we searched initially for homologs of each gene within the *E. coli* genome, which has the highest proportion of functionally annotated genes. The resulting network, involving only 10 of the 24 core genes, provided a very conservative view of the relationships and paralogy among the core genes but showed that flagellar genes were derived largely from other flagellar genes with apparently little input from other coding sequences.

All this means is that within a single *E. coli* genome (K-12) they didn't find much beyond the axial protein homologies.

> Extending these analyses to include other genomes uncovered additional links among flagellar proteins and revealed that the entire set of core genes could be formed through the duplication and divergence of previously existing flagellar genes.

Figure 3 again. Yep, I guess they really do think this. "Entire set" is not ambiguous.

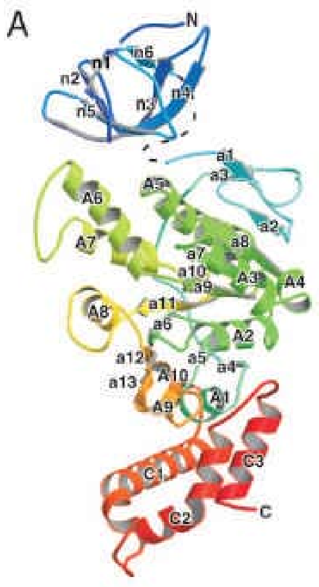
> That the analysis of the *E. coli* did not resolve all of the links among core genes is not surprising given that these genes are ancient and have followed independent histories within bacterial lineages. It was originally hypothesized that biological pathways and structures might expand through the successive addition and modification of their preceding components (27). Although there is diminishing evidence that the recruitment of new enzymes into metabolic pathways occurs by this process (28), it is apparently the manner by which the bacterial flagellum arose.

In other words, the all-internal-duplications model has failed for metabolic pathways, and the cooption-from-diverse-sources model is now dominant – but we'll resurrect the old discredited model for the flagellum!

> [end of the conclusion paragraph]
>
> As with the evolution of other complex structures and processes (29–32), we have shown the bacterial flagellum too originated from "so simple a beginning," in this case, a single gene that underwent successive duplications and subsequent diversification during the early evolution of Bacteria.

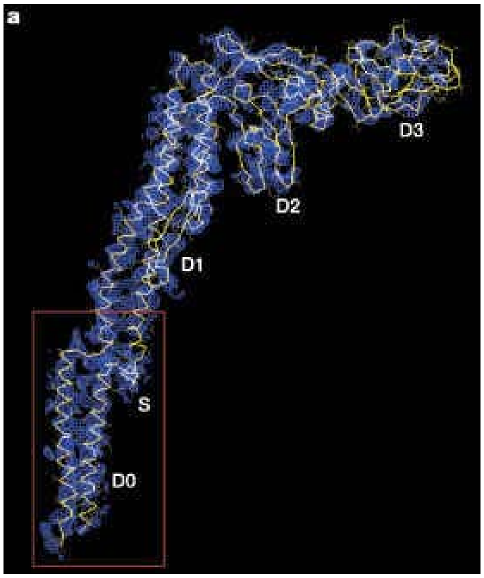
This conclusion sentence is the 5th or 6th time the authors explicitly endorse the "all flagellar genes came from one" model. So let's not have any more questions about what the authors meant to say.

Now, why am I skeptical? I posted some scattershot reasons on Monday. Some us PT posters are chewing on more technical issues. But to really get a visceral sense of the problem of claiming that all the flagellar genes are homologous, you just have to look at some structures. Here is FliI, the ATPase that powers the secretion of axial proteins like FliC:



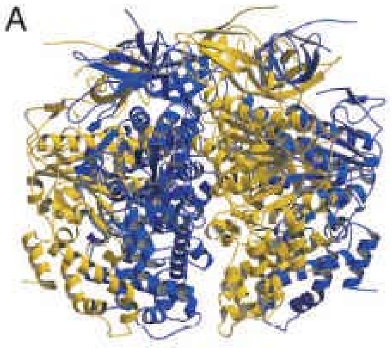
(from Imada et al. 2007, *PNAS*)

Here is FliC, aka flagellin, which is supposed to be just two homology links away from FliI in the Liu/Ochman Figure 3:



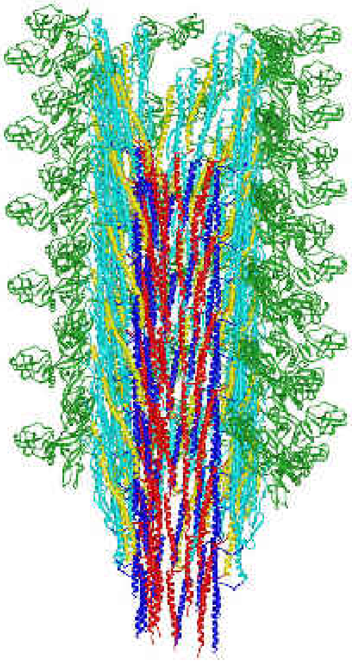
(from Yonekura et al. 2003, *Nature*)

You can also compare the two proteins when assembled into multimers. FliI assembles into a homohexamer:



(from Imada et al. 2007, *PNAS*)

FliC assembles into the flagellar filament:



(from here)

For comparison, this is what two protein structures look like when they are homologous. These are the homologous domains in CheC (a chemotaxis protein in some flagella), and FliM (part of the flagellum base, which interacts with chemotaxis proteins to produce the switching in rotation direction). (FliM is basically two main domains, homologous to CheC + FliN. Liu & Ochman caught the FliN homology but apparently not the CheC one.)

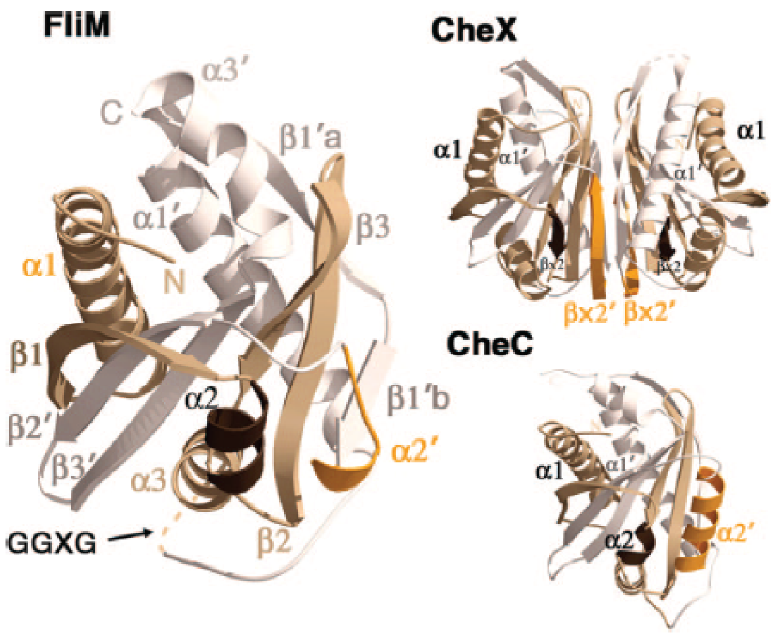


Fig. 2. Structure of FliM reveals homology to the CheC/CheX phosphatase family. Ribbon diagrams show topologies and secondary structural elements for FliM (*Left*), CheX (*Upper Right*), and CheC (*Lower Right*). Pseudo-2-fold axes relate one-half of the monomer units (white) to the other (tan). The $\alpha 2'/\beta_X'$ regions (orange), which differ in structure among the three proteins, dimerize CheX, associate CheC with CheD, and mediate FliM self-interactions. The conserved, but disordered, GGXG motif links the two halves of FliM.

(from Park et al. 2006, in PNAS)

You might have noticed that FliC and FliI do not exactly exhibit the kind of structural similarity shown by the homologous FliM and CheC.

I conclude that FliC and FliI are not homologous. Therefore not all core flagellar proteins are homologous. Therefore they didn’t come from a single ancestral ur-flagellar gene. Therefore the conclusion of Liu & Ochman is wrong. Case closed.

Commenters are responsible for the content of comments. The opinions expressed in articles, linked materials, and comments are not necessarily those of PandasThumb.org. See our full disclaimer.

Post a Comment

Use KwickXML formatting to markup your comments: <b>, <i>, <u> <s>, <quote author="...">, <url href="...">, etc. You may need to refresh before you will see your comment.

Name:

Comments:

Email Address:

URL:

Remember personal info?
○ Yes ● No

Preview Post
Check Spelling

Comment #171074

Posted by jv on April 20, 2007 3:58 AM (e)

I was in the “maybe there was a slight misunderstanding camp”, but I see that the flaws could go deeper. This is not, for the moment anyway, good work to use for ID bashing.

That’s okay for me, as I’m just a theorist – I’m not terribly worried about which specific mechanism it was; I just focus on the specifics of potential mechanisms. I’m happy to bash with the highly plausible existence of indirect mechanisms that they pretend don’t “really” count.

Comment #171075

Posted by Don Smith on April 20, 2007 4:03 AM (e)

A couple of minor edits:

I count 15 blue lines connecting the 10 yellow boxes.

> Here is FliC, aka flagellin, which is supposed to be just two homology links away from FliC in the Liu/Ochman Figure 3:

The second FliC should be FliI.

Comment #171082

Posted by Nigel Depledge on April 20, 2007 5:26 AM (e)

A very astute and incisive set of observations. Any chance of a bit of annotation on the structures, for those readers who don’t have experience in biochemistry? E.g. identification of beta-sheet and alpha-helices?

By the look of it, the FliC protein has loads of alpha-helix and no beta-sheet, which is rather different from the FliI protein. FliI seems to have several short alpha-helices, an anti-parallel beta-sheet, and maybe a little beta-strand.

Additionally, by my count, 18 of the homologies idenitfied in their figure 3 are supported by only 1 “hit” in 41 genomes, which is rather weak evidence on which to base such a sweeping conclusion.

Comment #171086

Posted by Daniel Morgan on April 20, 2007 6:03 AM (e)

We all know that protein structure and protein homology are notoriously difficult to correlate – elsewise, Pande wouldn’t have a gazillion PCs crunching the folding data. I, for one, would rather see the sequence data than the structural maps.

Comment #171092

Posted by TR Gregory on April 20, 2007 6:36 AM (e)

Hi Nick,

Thanks for the update – I look forward to seeing your reanalysis of their data and rebuttal of their interpretation in PNAS or another peer-reviewed venue.

Comment #171093

Posted by Ron Okimoto on April 20, 2007 6:52 AM (e)

PNAS is a fairly prestigious high profile journal to get published in, but back in the 1980's I got a chuckle out of one of the professors on my committee (frankly, I didn't notice him in the small departmental library) when I remarked to another grad student about passing the Proclamations of the National Academy.

I don't know what the review process is today, but back then members could publish their own work or submit the work of others for publication. To get published you would submit your research to a member and they would have it reviewed and if you passed it got published. There were leaks in the system. That is about all I have to say.

Comment #171100

Posted by Nick (Matzke) on April 20, 2007 8:52 AM (e)

Thanks – I fixed 14 and FliC/FliI mixup. Hard to count all them lines.

Comment #171109

Posted by Myrmecos on April 20, 2007 10:33 AM (e)

Nick,

I follow you through the problems with using BLAST to determine homology. That's a tricky subject and I think we all await the day when our homology searches are based more on model-based methods and less on a quick similarity algorithm.

But you lose me at the protein structure bit, where your assessment is little more than a subjective "These look different to me so they can't be homologous". That's not a rigourous argument, Nick, that's an assertive opinion, and it suffers from the same logical problem as claiming that horse's hooves and monkey's feet can't be homologous because they look different. You seem to be declaring that the external morphology of a molecule is more important in homology assessment than the sequence of amino acids within the molecule, and while that may or may not be true it is ad hoc reasoning to declare the one way is the correct way to do it.

In the meantime, what are we to make of the observation that there are no obvious homologies from some of the flagellar proteins to anything outside the flagellum? Your favored scenario implies co-option of proteins from other systems, but you seem locked in a logical spiral where your evidence for external homology comes from within the story that you wish to tell, not from any outside evidence of homology. This isn't to say that I've got anything against your scenario- I find it quite elegant- but your supporting evidence is at least as scant as that of Riu and Ochmann.

Comment #171112

Posted by Reed A. Cartwright on April 20, 2007 10:56 AM (e)

I think Nick's point is that the structures are very divergent, so much that existing models of flagella evolution do not propose that all flagellar proteins evolved from a common ancestor. The model of Liu and Ochman goes against current wisdom, and lacks strong evidence to back it up. Finding BLAST hits in a large database does not guarantee that those hits will be evolutionary significant. I haven't had time to look at the supplementary material, but several people I trust say that some of these homologies are artifacts.

Comment #171113

Posted by Myrmecos on April 20, 2007 11:06 AM (e)

Reed,

Having an internally consistent model for the evolution of the flagellum is not the same thing as having external corroborating evidence for that model. The rejection of putative homologies on the grounds that they don't fit with the model is precisely backwards for how one ought to do science, yet that is what you and Nick are suggesting. Shall we be critical of all data that don't fit Nick's scenario, or shall we be critical of Nick's scenario when the data don't appear to fit it?

There are certainly issues with using BLAST to assess homology, and perhaps those criticisms are well-founded, in which case my point is moot. But I'm not about to accept a blogosphere debate on the topic over the normal peer-review process, however flawed it may be.

Comment #171118

Posted by Bufo bufo on April 20, 2007 11:56 AM (e)

Hi Nick,

I'm no biologist, but this topic is interesting to me (and, I assume, many other non-experts), and I would like to understand your critique in more detail. As others in this thread have pointed out, your dismissal of the proposed homologies isn't really argued for in detail, and it's hard for someone with no experience in "reading" protein structures to understand the significance of the images you present.

I take it from your categorical "case closed" that you find the structures conclusive. Could you help those of us who want to understand the issue but don't have the necessary training by pointing out some of the features that lead you to this conclusion? Are there particular properties that rule out homology, or are the structures simply so different that any biologist would immediately (and correctly) decide that there is no chance that the proteins are homologous?

Also, I'm wondering whether there are any structure-based similarity measurements that could be used to quantify the similarity — somehow "they look different to me, and I'm an expert" strikes me as, well, unscientific..

Comment #171119

Posted by Reed A. Cartwright on April 20, 2007 11:57 AM (e)

Myrmecos,

You have it wrong. The paper is raising issues because it lacks convincing evidence for its novel homologies. A single BLAST hit out of what—over 100,000 comparisons—does not suggest that the hit is not an artifact. The gray lines are the blue lines don't seem to be comparable.

There is much more discussion going on in email than you have seen on PT. It is my understanding that those details will be published on PT.

Comment #171120

Posted by Pete Dunkelberg on April 20, 2007 11:59 AM (e)

"There are certainly issues with using BLAST to assess homology, and perhaps those criticisms are well-founded, …"

You could say that indeed, and that's what is being said here. Nick is not the only one looking at this. In fact he is communicating with others who are looking closely at the paper, and under the circumstances many more people must be examining it. Have you noticed any rebuttal?

“ … in which case my point is moot. But I’m not about to accept a blogosphere debate on the topic over the normal peer-review process, however flawed it may be.”

There will surely be more flagellum papers. Meanwhile I predict that the blog approach will turn out to be a constructive lesson, or lessons, for many people including some who don’t, and some who do, read the journals. You may like it better in time. I agree though that this was a bit startling, and different reactions are not surprising.

Comment #171122

Posted by David Stanton on April 20, 2007 12:21 PM (e)

Nick,

It seems the topic is too complex for a mere letter to the editor. It might be more appropriate to write a rebuttal paper. What you need is statistically significant evidence for homology of some of the genes to genes other than those proposed in the paper. Even just a few obvious examples should suffice. I still think it would be best to get some well known names in the field as co-authors.

Pete is right, we should learn some valuable lessons here about the role of the internet in such debates. Once this issue is resolved, it should serve as a shining example of the self-correcting nature of science. The most important thing to remember is that ‘poof’ is no longer on the table.

Comment #171123

Posted by Dan Gaston on April 20, 2007 12:21 PM (e)

> ***Daniel Morgan wrote:***
>
> We all know that protein structure and protein homology are notoriously difficult to correlate – elsewise, Pande wouldn’t have a gazillion PCs crunching the folding data. I, for one, would rather see the sequence data than the structural maps

It isn’t so much that protein structural homology and protein sequence homology are hard to correlate. The protein folding problem is usually approaching it from an ab initio approach (and this is slightly off topic but I work with protein sequence/structure homology quite extensively). which is why it takes so much computational power. Homology modeling on the other hand, which relies explicitly on the correlation between sequence homology and structural homology (because structural homology is retained much longer than sequence homology) is quite robust.

However I too would like to see the blast results themselves and the pairwise alignments generated. Their scores are fairly significant for detecting homology but artifacts do occur. When I was reading the paper I would have been fairly comfortable with a declaration that the flagellum arose through the successive duplication of several ancestral proteins and I think their data would have supported that position.

Comment #171124

Posted by David vun Kannon on April 20, 2007 12:25 PM (e)

Nick, I agree with your conclusion on what Liu and Ochman are asserting. I side with several other commenters that your argument from structure is weaker than it could be.
I know the tertiary structure is notoriously difficult to derive straight from the genetic sequence. Your argument seems to be that all homologies in sequence space should be obvious from arbitrary rotations of 3D structures. I can’t buy it.
Park, et al. 2006 wants to convince us of the homology of FliM and CheC, and rotates and colors the model to help make that argument. You take two proteins that not homologous to each other, but rather to an intermediate,

which are modeled and rotated arbitrarily with respect to each other, using different graphical systems, and expect the reader to conclude based on appearance that you have invalidated Liu and Ochman’s argument.

Sorry, you lose. If I squint hard enough, I can see the tail of FliC as an extension of the lower part of FliI, and the assembled FliC as a stacked set of homo-n-mers.

Your argument would be more convincing, though not dispositive, if you could show the intermediate protein’s teriary structure, show all three as ribbon diagrams of the same orientation and coloring, and if you could show that the changes in sequence should not be expected to result in the corresponding changes in tertiary structure. My naive expectation is that the intermediate has a tail of size intermediate between FliC and FliI. You’d go a long way to convincing me if you could show my expectation is wrong.

To the point that some of the lines are supported by only one hit out of 41, I would counter that (as you point out) there is some conservatism in e value. A looser e value might support more hits, but the fact that the authors stuck with a conservative value argues in favor of an attempt to do careful work and following the data to a conclusion, rather than a BLAST fishing expedition.

Comment #171126

Posted by Nick (Matzke) on April 20, 2007 1:02 PM (e)

Cripes, how much different could those structures get? At the very least it would take a detailed argument to conclude homology, not 2 hits out of 41 genomes (which is the only thing that connects FliI to FliD, which is connected to FliC).

Typically the rule is:

Sequence similarity that is weak, but statistically strong enough to conclude homology, predicts almost superimposable structures in the homologous domains, e.g. like FliM and CheC.

Here’s another argument for people. FliI is an ATPase. It is highly conserved, with well-known homologous in – literally – hundreds of different relatives. The closest relatives in BLAST and PSI-BLAST searches go something like this: FliI, F1-ATPase, RecA, PilT (from Type 4 Secretion), then the whole superfamily of AAA ATPases, and then even more. Heck, FliI is more closely related to *eukaryotic dynein* than it is to FliC or FliD. At least FliI and dynein both form similar hexameric superstructures that use ATP (with sequence-homologous domains) to effect conformation change, and at least this hit can be retrieved on PSI-BLAST (many of the Liu-Ochman claims cannot…we have tried).

For ATPase homologs see: Iyer LM, Leipe DD, Koonin EV, Aravind L. “Evolutionary history and higher order classification of AAA+ ATPases.” *J Struct Biol*. 2004 Apr-May;146(1-2):11-31.

Comment #171146

Posted by David Stanton on April 20, 2007 2:36 PM (e)

Nick,

Have you considered a phylogenetic analysis of the appropriate genes? The homology hits and conserved structure approaches are not really the same thing as a cladistic analysis. Perhaps showing that some of the genes were more closely related to other genes than the ones proposed in the paper would help the argument. You could determine the degree of confidence in the tree and determine how many more steps it would take to force the topology into that proposed in the paper. I’m not sure how difficult this would be, since alignment issues might complicate things. Still, if you have sequences for the candidate ancestral genes it might be worth a try.

Comment #171164

Posted by David vun Kannon on April 20, 2007 4:44 PM (e)

> ***Nick wrote:***
>
> Cripes, how much different could those structures get? At the very least it would take a

> detailed argument to conclude homology, not 2 hits out of 41 genomes (which is the only thing that connects FliI to FliD, which is connected to FliC).

Do I hear an argument from incredulity? I'm not clear why you think it would take a detailed argument to prove homology, when a handwaving argument was enough to disprove it in your post.

If Liu and Ochman have built a cladogram that depends upon shaky homologies in your opinion, just say so. You start off by calling their pairwise homology hits "significant", so I think you've undercut that argument. If they had published Fig 3 with e=0.01, gotten only 2 out of 41 pairwise homologies, and then made strong claims, you'd be home free. But at e=0.0001 it looks like you've got to work a little harder to convince me.

Comment #171168

Posted by Michael Buratovich on April 20, 2007 5:08 PM (e)

Also let's remember that proteins can have similar three-dimensional structures and still have little to no primary sequence similarity. For example mandelate racemase and muconate lactonizing enzyme from Pseudomonas ovalis have virtually identical 3-D structures even though they only share 26% sequence homology. L-aspartate amino transferase and D-amino acid aminotransferase (PDB 2aat, 1daa) have similar tertiary structures even though they have very little in common, sequence-wise. Likewise tryptophan synthase (PDB 1ttp) and galactonate dehydratase have very similar alpha/beta barrel domains even though there is no sequence homology between the two proteins.

Comment #171172

Posted by Reed A. Cartwright on April 20, 2007 5:16 PM (e)

David,

A significant hit in a BLAST search is not the same as a significant relationship evolutionarily. Unrelated proteins can have parts that are similar due to convergent evolution.

Comment #171173

Posted by Torbjörn Larsson on April 20, 2007 5:17 PM (e)

> ***Nick Matzke wrote:***
>
> Because (a) is a higher proportion than (b), the authors say that "on average" the within-flagellum homologies are more common. OK I guess. This doesn't establish anything about all the flagellar proteins evolving via internal duplications – if you've got external homologies then you have evidence that flagellar proteins could have once had an nonflagellar ancestors.

This poor non-biologist can't follow far into the wondrous land of structural vs sequence homology. But this suggestion that the grouping may be an artifact of the analysis seems easy pickings.

> ***Nick Matzke wrote:***
>
> Here's another argument for people. FliI is an ATPase. It is highly conserved, with well-known homologous in – literally – hundreds of different relatives.

There you go.

Comment #171193

Posted by Clastito on April 20, 2007 7:43 PM (e)

I can believe, you NIck. PNAS has been publishing lots of really crappy stuff.

Comment #171194

Posted by bdelloid on April 20, 2007 7:52 PM (e)

The use of other genomes to identify related sequences is entirely reasonable - for the same reason, it is helpful to have “dense taxon sampling” when resolving phylogenetic relationships. Furthermore, distantly related sequences can help resolve taxonomic relationships between closely related species. This is a fact.

Nick, if a single lineage within E. coli is fast evolving, but more slowly evolving in another species, then it is perfectly reasonable to ID relatives by BLAST in other species that are not found in the same genome.

Furthermore, as others have commented, the structure homology assertions have no logical extra weight than sequence homology assertions. This is the sad truth about comparing distantly related sequences

I have to say you are completely wrong here.

Comment #171195

Posted by bdelloid on April 20, 2007 8:10 PM (e)

Please sir, may I kindly backpedal? Somehow I got carried away with an argument that I thought you were making, but weren’t.

My bad.

Comment #171197

Posted by Dan Gaston on April 20, 2007 8:34 PM (e)

I also feel compelled to point out that the second author and the editor of the paper are very, very good scientists who publish some damn good work. I’ll admit I should go back over the paper a little more critically, and I don’t agree with the assertion that there was one ancestral sequence as opposed to a small group of ancestral sequences. There may be something off, flagellum evolutionary relationships isn’t exactly my expertise, but I’m not ready to hang the authors out to dry just yet like some seem to want to do. Critiquing the paper is all well and good but if in my opinion if you disagree vehemently you work up a paper of your own and publish a rebuttal, or you write something as a commentary.

Comment #171201

Posted by David Stanton on April 20, 2007 9:15 PM (e)

“A significant hit in a BLAST search is not the same as a significant relationship evolutionarily. Unrelated proteins can have parts that are similar due to convergent evolution.”

Reed,

Thanks for responding. That is exactly the point I was trying to make. This is why BLAST searches can yield spurious results. The figures from the paper are not really a phylogenetic analysis.

What I had in mind was something like what Morris Goodman did for the globin gene story. All the genes were derived from one ancestral sequence. He demonstrated their relationships and estimated divergence dates with a phylogenetic analysis. If any of the genes were not derived by duplication from the ancestral sequence, they would not fit into the same clade as the other genes. If you included other possible ancestral genes in the analysis, the genes derived from them should group with them rather than with the other clade.

Of course this problem is a little more complex. The divergence dates are longer. There are bound to be problems with alignment, long branch attraction, identification of ancestral sequences, etc. Still it might be worth a try.

Sorry if I'm being too simplistic or if there is something I'm missing. Just a suggestion.

Comment #171203

Posted by Reed A. Cartwright on April 20, 2007 9:23 PM (e)

David Stanton,

Note that my comment was directed to the other David.

Comment #171208

Posted by David Stanton on April 20, 2007 9:45 PM (e)

Reed,

Sorry.
What's all this about Presidential Aids?
Never mind.

Comment #171210

Posted by bdelloid on April 20, 2007 9:55 PM (e)

The idea that a phylogenetic analysis somehow gives information about homology but a BLAST hit doesn't isn't correct. If I can find a protein by BLAST, then I can align those proteins. If I can align those proteins, I can build a tree. By building a tree one is already, by definition, assuming homology.

Now, I will agree that a phylogenetic tree can give information about degrees of relatedness. But the definition of homologous is that the sequences are "descendants of a common ancestor." The only way two proteins could not be homologous at some level is if at least one of them evolved de novo out of dna sequence that was not derived in any way from the ancestor of the other protein. Like an intron evolving into an exon...

Comment #171216

Posted by David Stanton on April 20, 2007 10:51 PM (e)

Bdelloid,

That is exactly the issue here. Were all of the genes derived from a single common ancestor, or were they derived from separate ancestral sequences. It seems to me that a BLAST search is only the first step in answering the question.

Of course all genes are related in some way. The question is, what genes shared a common ancestral sequence more recently and which are not part of that lineage. It seems to me that a phylogenetic analysis would be helpful in addressing this issue. Of course that's only my opinion. I could be wrong.

By the way, I really like your handle. One of my favorite groups.

Comment #171224

Posted by bdelloid on April 20, 2007 11:32 PM (e)

David,

Again, I've thought about my comment more after posting it. You are right, the phylogenetic analysis would be helpful to answer whether the units of the flagellum are related, to the exclusion of other proteins, as a clade with a single common ancestor.

Still, at some level it seems likely that they are descendants of a single common gene - I guess the question is whether the units of the flagellum are paraphyletic or monophyletic.

http://en.wikipedia.org/wiki/Paraphyletic
http://en.wikipedia.org/wiki/Monophyly

Comment #171227

Posted by David Stanton on April 20, 2007 11:57 PM (e)

Bdelloid,

Thanks. But after further reflection I have to admit you were right! When you align sequences it is an hypothesis of homology, so aligning genes that are really not homologous would not work.

I guess the best you could do with this approach would be to show that some of the genes were a kind of out group to the clade consisting of the genes that were descendants of the same ancestral lineage.

Ironically then, it seems that this approach would work,(in much the same way it works with ribosomal and hemoglobin genes), but only if Nick is wrong and the authors of the paper are right. Then the genes could be reasonably aligned and a phylogenetic analysis would be appropriate. Oh well, I guess if it were an easy question to address it would have been answered long ago.

Regardless, it is a pleasure discussing things in a civil manner with a reasonable person.

Comment #171230

Posted by bdelloid on April 21, 2007 12:52 AM (e)

Cheers. I was able to stay on Uncommon Descent for a real long time by being civil, but my last round of comments didn’t show up. Oh well.

Great_ape is the real king of civil discussion - he’s been posting there forever.

Comment #171266

Posted by David vun Kannon on April 21, 2007 10:06 AM (e)

> ***Reed wrote:***
>
> A significant hit in a BLAST search is not the same as a significant relationship evolutionarily. Unrelated proteins can have parts that are similar due to convergent evolution.

I think I’m the other David you were responding to.

Yes, that’s a good argument. But it’s not the argument Nick made in his post. Nick accepted the quality of the BLAST hits, but argued from incredulity on the tertiary structure.

I’m sure someone could actually study the raw probability of that kind of convergent evolution happening, as well as the observed probability over a number of sequences, and come up with a null hypothesis to test these results against. But until that study appears it’s just a good, but handwaving, argument. Personally, I think the probability of convergent evolution over extended sequences is much much lower than 2/41. If convergent evolution of sequences was a significant concern, I would expect many cladistic analyses would have to be hedged in their conclusions. I’m no expert in the literature, but I don’t recall seeing that kind of careful hedging wrt convergence.

Comment #171267

Posted by David vun Kannon on April 21, 2007 10:16 AM (e)

BTW, I love the way Nick and PZ Meyers become reliable sources for Behe and other IDists to quote - when it suits their purpose! Where is their respect for Nick's acquired expertise in bacterial flagellum issues the other 99.99% of the time? It's not quote mining, it's reputation mining...

Comment #171272

Posted by Dan Gaston on April 21, 2007 11:13 AM (e)

In terms of blast, alignments, homology, and tree building I just want to make a few quick points that follows a sort of step-by=step sequence of rules of thumb for determining homology. As always when looking at homology convergent evolution is always a problem, and difficult to deal with. I'd also like to point out that using the inter-genomic BLAST within flagellated bacteria, as far as I know, is quite acceptable for the purposes the authors used it. This is how people doing bacterial genomic analysis frequently operate and it has proven to be a powerful tool. Anyway as for homology:

1) Is the BLAST hit significant? If Yes move on to step 2

2) Look at the pairwise alignment, how similar are the two sequences? How good is the alignment? If the percent identity is high we can usually be justified in declaring them to be homologs. When we get lower, down close to the twilight zone of sequence homology we can usually say the sequences are homologous but we should also do some other tests as well to be sure. Once we get down into the twilight zone homology is almost impossible to declare based on sequence, definitely check other methods such as structural alignments to test for homology.

3)The "other tests" of which there are a wide variety. Structural analysis/alignment, phylogenetic analysis and others. Do the sequences share a major domain with one another? That sort of thing. If they are not complete homologs are they partial homologs if that is the case?

When you are doing alignments you aren't explicitly assuming they are homologous, you are testing to see if they may be. If the alignment is good chances are that they are, convergent evolution on a large enough scale to produce a really good alignment over a whole sequence length is pretty rare after all. If you only produce an alignment over a domain than they may still be fully homologous because evolutionary pressures only constrained that domain in one lineage and not in the other, or they may be partial homologs.

When we perform a phylogenetic analysis yes, we are sort of assuming that the sequences are homologs, BUT we are also testing that assumption as well. Throwing non-homologous sequences can really screw up your tree. This isn't always going to be caught but it is one of the things we keep in mind when doing this sort of analysis.

Comment #171274

Posted by Nick (Matzke) on April 21, 2007 11:24 AM (e)

I guess I should have made it more explicit, although I alluded to it several times – the good e-values on the BLAST hits that the authors report are not reproduceable with their stated methods. It is simple enough to get bl2seq from the NCBI website and check this. Thus there is substantial reason to be skeptical of the claim of all these hits below e=0.0001 also – which was obvious enough on Monday, because e.g. for Pallen/Matzke 2006 we BLASTed everything, on the entire non-redundant sequence database not just 41 genomes, and didn't get most of those hits.

I agree that just eyeballing structures is not hugely rigorous and there might occasionally be ways that structures diverge while sequence similarities remain, but (a) usually structure is more conserved than sequence, (b) the null assumption is always no homology, you would need to make an extensive argument, present alignments for inspection, etc., to connect FliI to FliC, and explain why the structural differences were not contradictory information, and © those of you who see vague similarities between FliI and FliC, remember that you also have to connect the FliM structure *also*, plus several proteins that are nothing but transmembrane helices in the inner membrane, etc.

Comment #171280

Posted by David Stanton on April 21, 2007 11:58 AM (e)

I mentioned that a BLAST search was only the first step in such an analysis. Thanks to Dan for his excellent clarification of what some of the next steps might be.

It seems that there could be significant difficulty in sequence alignment for this problem. Perhaps there might be some other “higher order” genetic features that might be used to address the issue of “monophyly” of this gene lineage. Perhaps intron structure, promotor or enhancer sequences, or signaling sequences could be examined to try to determine the relationships between these and other genes. Who knows, maybe even gene mapping studies might be useful.

It also seems that there are difficulties with structural comparisons as well. Still, they will undoubtedly play a significant role in illucidating the “final” answer. Good luck to Nick in this most difficult task.

Comment #171282

Posted by jkc on April 21, 2007 12:44 PM (e)

I have a really dumb question, but since I am not a scientist and a relative newcomer to this topic, I will ask it anyway: Is this line of research really that intrinsically interesting, or did it become more interesting after Behe published his infamous book?

Comment #171283

Posted by Jim Hu on April 21, 2007 12:53 PM (e)

My early thoughts, just on the structural biology aspect of the argument here. Much shorter version: While structural similarity provides evidence *for* common descent, lack of structural similarity may not mean lack of common descent when looking at deep divergence times.

Still have to digest the paper though. I’m not convinced that FliI and FliC *are* homologs. It does look like they presented enough detail to allow others to reproduce their analysis in sufficient detail to examine things they didn’t have room to present.

Comment #171287

Posted by Michael Buratovich on April 21, 2007 1:40 PM (e)

jkc,

I can only speak for myself. Homology and gene lineages are BLOODY interesting whether or not Behe published is book.

Comment #171296

Posted by Pete Dunkelberg on April 21, 2007 2:26 PM (e)

It’s really interesting and a standard, important sort of research. The creo side show is irrelevant, except to politics. I guess most folks have no idea how much research goes on routinely.

Comment #171298

Posted by David Stanton on April 21, 2007 2:33 PM (e)

JKC,

I my opinion, yes. The illucidation of the origin of novel morphological features has been one of the most fascinating aspects of evolutionary biology since the time of Darwin.

I would think that the study of the origin of eyes, limbs, bird feathers, insect wings etc. should be intrinsically interesting to most biologists. Behe just chose a technically difficult question because he probably hoped that it would be a long time before any real answers would be forthcoming on this particular issue.

Unfortunately for Behe, advances in DNA sequencing and comparative genomics have gotten to the point where we now have the tools needed to start to address issues such as this. By the way, lots of good biologists were working on this particular issue long before Behe ever put in his two cents worth. Behe is certainly not the only reason there is interest in this topic.

Comment #171303

Posted by Dan Gaston on April 21, 2007 3:15 PM (e)

Nick,

If their results are not reproducible that is one thing but I think you may be missing a key point. BLAST e-values are dependent on the size of the database queried because it is a representation on the number of times you would expect to find a sequence from your database to align with your query sequence and have the exact same raw score. Unless I missed something they are NOT blasting against NR they are using their own databases. One dataset is blasting against *E. coli* only and then they extend that to their 41 complete genomes. And there is nothing wrong with that experimental procedure, it is quite common in doing comparative genomics/evolution in bacterial species.

In order to reproduce their work and those e-values you have to use the same datasets that they used. I'm not arguing for or against the paper here by the way, I'm just pointing out some things I have seen said so far that are not 100% correct. If the 41 genomes they used are publicly available then you want to start by downloading those, installing BLAST locally, and then you can start checking their methods for problems.

> ***jkc wrote:***
>
> I have a really dumb question, but since I am not a scientist and a relative newcomer to this topic, I will ask it anyway: Is this line of research really that intrinsically interesting, or did it become more interesting after Behe published his infamous book?

It's always been pretty interesting, at least for those interested in evolution and in particular interested in the evolution of complex molecular machinery. I'm more interested in the evolution of protein structure itself at the level of individual proteins but this is still pretty cool stuff.

> ***David Stanton wrote:***
>
> It seems that there could be significant difficulty in sequence alignment for this problem. Perhaps there might be some other "higher order" genetic features that might be used to address the issue of "monophyly" of this gene lineage. Perhaps intron structure, promotor or enhancer sequences, or signaling sequences could be examined to try to determine the relationships between these and other genes. Who knows, maybe even gene mapping studies might be useful.

Well in this case there are no intronic sequences since we are dealing exclusively with bacteria and the only signal sequences are for export from the cell I believe (if I remember all my bacterial genetics properly.) The good thing is that in that regard bacterial sequences are easier to work with, the difficulty is things like rampant lateral gene transfer, co-option, etc which can complicate this sort of analysis.

Comment #171315

Posted by Nick (Matzke) on April 21, 2007 5:26 PM (e)

Yeah, we are aware of the database size issue, it is one of several things in the technical category we are looking at.

Comment #171318

Posted by jkc on April 21, 2007 5:49 PM (e)

Michael, Pete, David, Dan (& Nick),

Thank you for responding (and also not assuming that I was trying to start trouble). I am gratified to hear that this is a fruitful line of research in its own right, and not just a wild goose chase necessitated by a desire to respond to misguided creationist scientists. My son wants to pursue a career that combines statistics (my field) and biology. Maybe I will nudge him in the direction of bioinformatics.

Comment #171347

Posted by Pete Dunkelberg on April 22, 2007 12:38 AM (e)

> My son wants to pursue a career that combines statistics (my field) and biology. Maybe I will nudge him in the direction of bioinformatics.

A strong combination. Doubtless you are aware of R A Fisher and know that fundamental statistical work was inspired by biological questions and lays at the basis of the modern synthesis.

Having bioinformatics in his tool kit is one thing. Check _Bioinformatics for Dummies_.
But there's lots to think about and lots of biology to learn before specializing in bioinformatics as such.
A look at journals like Evolution, Ecology, and American Naturalist, as well as basic texts on population genetics will provide plenty of ideas. Landscape Ecology [recent Conference] is an emerging field of increasing importance as human population increases and global climate changes .

Comment #171364

Posted by Luna_the_cat on April 22, 2007 8:39 AM (e)

This is disappointing. Maybe I'm off-base here, but it seems to me that doing a sequence homology search on some of these becomes largely pointless where you have to set your parameters so loose in order to catch true positives that your response set becomes flooded with false positives. Suspected homology HAS to be compared using structural homologies as well, surely, to correlate with structural conservation and weed out noise. And although I've only skimmed this paper, it kind of looks to me like they are not weeding out noise, and are engaging in a lot of wishful thinking about what might be a legitimate paralogy as opposed to what might have that small percent similarity through sheer coincidence.

So, has anyone actually done a node-and-edge graph to look at possible conserved structure?

Comment #171408

Posted by Dan Gaston on April 22, 2007 3:00 PM (e)

JKC

I'm currently working on a Masters in Computational Biology and Bioinformatics, which of course is a broad discipline in its own right. I happen to work in a molecular evolution research group, and there is a lot of stats at play there as well. I think if your son wants to combine Stats and Biology than Computational Biology is a good start but also pick an area of research that you are interested in within biology. While I enjoy the computational side of things it wouldn't be as interesting if the biological problem didn't also get the blood going.

I don't think I would enjoy it nearly as much if I was just working with computational issues with micro array data for instance (no offense to those who do that kind of work, it just isn't my cup of tea :))

> ***Luna_the_cat wrote:***

> This is disappointing. Maybe I'm off-base here, but it seems to me that doing a sequence homology search on some of these becomes largely pointless where you have to set your parameters so loose in order to catch true positives that your response set becomes flooded with false positives. Suspected homology HAS to be compared using structural homologies as well, surely, to correlate with structural conservation and weed out noise. And although I've only skimmed this paper, it kind of looks to me like they are not weeding out noise, and are engaging in a lot of wishful thinking about what might be a legitimate paralogy as opposed to what might have that small percent similarity through sheer coincidence.
>
> So, has anyone actually done a node-and-edge graph to look at possible conserved structure?

Structural conservation is generally retained longer than sequence identity it is true, and tends to be more well conserved. But inferences of homology can be made without structural comparisons depending on the quality of the alignment and the scores involved, often times we don't have good structures to work with after all. Convergent evolution of course is always a problem to worry about, and that holds true for sequences and for structures, some would argue that on certain structural levels it is even more of a concern because structural space is most definitely much smaller than the allowable sequence space in terms of proteins.

In terms of what they did, their methodology is fairly routine these days in bacterial studies of this nature. Limiting your database to the complete genomes of interest and doing comparative sequence analysis. Whether you think their data supports their conclusion is one thing. I think some people have possibly jumped the gun and started slinging around phrases like :serious methodological problems" far too early. That is a pretty serious thing to toss out there in scientific circles and remember these are some good scientists we are talking about, and the editor in charge of the paper is also very good so I am sure the reviewers for the paper were excellent choices as well. If there are problems it will be shown, I applaud people for reading critically and all of that. But some of the response has been a little too pointed and with an accusatory tone to it that I have found slightly disrespectful and perhaps premature.

Comment #171442

Posted by jjj on April 23, 2007 2:53 AM (e)

Nick calm down. Your argument is terrible. You download two pdbs, take a look at them from different angles, and say that the paper is wrong because the structures don't look similar to you?

Here is another argument I like: you did the BLAST searches yourself and – oh my gosh – the e-value doesn't match up! Looks like these authors forged some data! Then Dan comes along and tells you that e-value depends on database size and you reply that you *already knew that.* Well, if you already knew that, then why would you try to use it to discredit this paper?

I can't figure out what makes you think that two homologous genes need have similar protein structures. To help you understand this, why don't we imagine the following scenario. We start with a gene, a frameshift mutation occurs, and we end up with a different gene with a new function. We only have ONE mutation in this gene and the two genes are obviously homologous, but since it is a frameshift, most of the amino acids have changed and we have a completely different structure. I am not saying that this is what happened with the flagellar genes – simply that homologous genes need not have similar protein structures.

You have your agenda. Why don't you stick to the science?

Comment #171444

Posted by lllll on April 23, 2007 3:18 AM (e)

Also, since you have your pdb viewer up and running, why don't you take a look at what happens to protein structure when you mutate, e.g. prolines to alanines. Such mutations may have a huge impact on protein structure – of course, they are normally selected against, but at some point in time we have to accept the fact that improbable events occur (in other words, beneficial mutations DO happen).

My point is that you cannot discount homology – especially when there is strong evidence SUPPORTING homology in the paper – on the grounds that the proteins do not look similar to you.

Comment #171447

Posted by Ralf Koebnik on April 23, 2007 3:43 AM (e)

Hi,

concerning the question whether two proteins which are similar in sequence are also similar in structure - or vice versa - I would like to draw your attention to this pretty instructive paper:

Rost B.
Twilight zone of protein sequence alignments.
Protein Eng. 1999 Feb;12(2):85-94.
PMID: 10195279 [PubMed - indexed for MEDLINE]

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?
cmd=Retrieve&db=PubMed&list_uids=10195279&dopt=Abstract

From this paper one can conclude that proteins which are similar in structure are usually not similar in sequence (as judged, e.g., by BLAST) but that two proteins of a certain length and degree of sequence similarity are probably also similar in structure (for details, see the paper).

Hence, absence of structural similarity is a better indicator than absence of sequence similarity to show that two proteins are not related to each other, i.e. that they are not homologous. May be this is what Nick had in his mind when he put the 3D comparisons as arguments.

Cheers, Ralf

Comment #171448

Posted by Nick (Matzke) on April 23, 2007 3:45 AM (e)

It's not just me…

Comment #171465

Posted by Dan Gaston on April 23, 2007 7:25 AM (e)

Ralph:

Thats very true, which is what I was referencing when I was saying that structural similarity usually stays while sequence similarity may have been lost., such as in the "twilight zone". It is pretty rare for sequence mutation apparently to result in rather dramatic structural changes and not to have been weeded out because of purifying selection although one has to be careful when assessing structural homology. Not sure if anyone else here has done structural alignments but let me just say that the techniques are not as easy or in my opinion quite as refined yet as with sequence alignment. There are some good methods but we are dealing with crystal structures (or NMR) and even two clearly homologous protein structures can be difficult to align.

If anyone wants to check two protein structures I'd suggest Combinatorial Extension or FATCAT. The benefit of FATCAT is that it allows twists in the structures which is reasonable and can result in huge improvements in RMSD scores.

Nick I know many people are unhappy with this paper, and I'm not exactly trying to come down on one side or the other. I've just been trying to point out a few things that I think may have needed clarification for some of the readers here as well as pointing out that some of the tone taken against the authors just seemed a little to accusatory and strident for what should be civil, rational, scientific discourse. Thats just my opinion anyway.

Anyway if you are unhappy with their BLAST results just be sure you are working with the same dataset and using the same method. blast2seq at NCBI (or any BLAST tool at NCBI for that matter) is going to be completely useless in terms of comparing e-values. You'll have to have the 41 genomes set up as a local database with a local install of BLAST. Its really the only way you can do a baseline comparison and validate results.

Comment #171518

Posted by Douglas Theobald on April 23, 2007 12:53 PM (e)

> ***Dan Gaston wrote:***
>
> Anyway if you are unhappy with their BLAST results just be sure you are working with the same dataset and using the same method. blast2seq at NCBI (or any BLAST tool at NCBI for that matter) is going to be completely useless in terms of comparing e-values. You'll have to have the 41 genomes set up as a local database with a local install of BLAST. Its really the only way you can do a baseline comparison and validate results.

Not if the authors didn't correct for the size of the database ...

Comment #171521

Posted by Reed A. Cartwright on April 23, 2007 1:13 PM (e)

> ***jjj/lllll wrote:***
>
> My point is that you cannot discount homology – especially when there is strong evidence SUPPORTING homology in the paper – on the grounds that the proteins do not look similar to you.

Except that there isn't strong evidence supporting the homology of FliC and FliI in the paper: no alignments, no phylogenies, not even BLAST scores. They only evidence they offer is that in two genomes bl2seq found some hit between FliD and FliI at a cutoff of 1e-4. That is not all that convincing, especially given the large number of comparisons that they did.

The novel and most interesting thing about this paper, the argument that all core flagellar genes descended from one original ur-flagellar gene is simply not well supported in the results provided by the paper.

PS: For full disclosure, I went to grad school with the lead author; we spent time in the same lab together. I also interviewed for grad school with the second author.

Comment #171541

Posted by Dan Gaston on April 23, 2007 2:47 PM (e)

> ***Douglas Theobald wrote:***
>
> Not if the authors didn't correct for the size of the database

You don't do the e-value calculation yourself, it is output by the BLAST program and is calculated internally based on the database you are BLASTing against. There isn't anything for the authors to adjust.

Comment #171552

Posted by Douglas Theobald on April 23, 2007 3:12 PM (e)

> ***Dan Gaston wrote:***
>
> You don't do the e-value calculation yourself, it is output by the BLAST program and is calculated internally based on the database you are BLASTing against. There isn't anything for the authors to adjust.

Sometimes there certainly is. If you use bl2seq with defaults, it gives you the e-value using one sequence for the query length, and the other sequence as the database length. Bl2seq compares one sequence to one other. If you use bl2seq, say, 1000 times, you **must** correct those e-values because they were not calculated by bl2seq using the entire database of 1000 sequences that you actually searched against. Of course regular BLAST, e.g. when using the standalone blastall executable, does this correction inherently, because it knows the database length; bl2seq, however, cannot in principle know the database length, other than the individual sequences you give it. In fact, bl2seq has the '-z' option so that you can do just this, you can supply the real database length so it can calcualte the proper e-values.

So, as I said, you can replicate the authors' results easily with bl2seq if they didn't do the proper correction – oh perhaps overestimating the significance of their hits by three orders of magnitude :).

Comment #171557

Posted by Douglas Theobald on April 23, 2007 3:32 PM (e)

Sorry, with bl2seq that should be the '-d' flag. The '-z' flag is the equivalent for blastall.

Comment #171572

Posted by Dan Gaston on April 23, 2007 4:37 PM (e)

Douglas:

I was under the impression they were most likely using blastall as opposed to blast2seq for their studies although I could be mistaken. If they used bl2seq I'm sure that they would have adjusted accordingly. I don't generally use bl2seq myself as I'm usually doing larger scale analysis.

Comment #171573

Posted by Dan Gaston on April 23, 2007 4:42 PM (e)

Hahaha nevermind, I'm a complete idiot. Of course they used Bl2seq for their pairwise comparisons. So yes the point stands that e-value correction is required here. Thanks for reminding me **Douglas**

Comment #171640

Posted by Nick (Matzke) on April 24, 2007 2:40 AM (e)

Here is what we mean by methodological problems and database size issues.

Trackback: Flagellum evolution kerfluffle continued

Posted by The Panda's Thumb on April 24, 2007 2:41 AM

As the discussion over the Liu-Ochman flagellum evolution paper continues, it is clear that I need to do a little more arguing to defend my position. Although some were convinced that skepticism was justified based the previous PT posts...