StORe

Interview Transcript

Interview reference: Biochemistry 10
Role: Post doc researchers from the same lab group
Interview length: 1hr 13mins
Format: Face to face
Number of interviewees: 2
Questionnaire respondent? No

A. Identities

A1. Which discipline group best describes your field of interest?

Biochemistry.

A2. Briefly describe a typical research project workflow. What steps do you go through e.g. from generating data, then analysing it and eventually publishing?

Interviewee 1
I'm working on structural genomics. I'm the guy that does target selection for our consortia. So for example, the starting point for the last project was the need to identify target families in order to characterize structures within the family. The first thing to do is to identify all the protein families and then categorize these according to criteria. e.g. how many of these families have a particular protein structure – we're going for a coarse-grained approach.

The workflow then is first to identify target families, so you need to identify a resource that describes them. We use a resource called PFAM and our own database CATH, which a lot of our group is based around. CATH families have structures – it's a structural database - and quite a few PFAM families - it's a sequence database - have structures too, so you can compare the two resources and then identify all the PFAM families that don't have a structure. So then you can say, we want to solve the structures for these families, so if you have the structure of a large family you can cover lots of characteristics of that structure. The next thing in terms of workflow is to try to estimate the probability of solving the structure in that family; can you express the protein?

Once you've identified all these things you need to prioritize them. We have four groups working together so a lot of our data is presented on the web so that we can compare each other's findings and ideas. We download each other's lists of data, run our own selection criteria on them and put them back in the pot. Different people may reject families for different reasons but you want to come up with an overall set of families that we all agree on because in the end the families will be distributed throughout the consortia. So there's a lot of interaction that goes on and the web has been really useful.

Interviewee 2
I work on various bioinformatics projects, mostly looking at database integration.

I have two jobs really: possibly the more interesting is the maintenance of the CATH database. In the last year or so we've been redeveloping the workflow prior to taking structures and putting them in the database. This starts with us getting new structures from wwPDB. We compare these with structures already in the database and if they're similar enough, then we can apply previous findings to these structures, break them up into protein domains and classify them.

If we can't do that then we go to manual stages: first someone chops up the structures using web pages to identify where within the sequence you can define the domain boundaries. Once it's chopped into domains you have another manual process of identifying where the domains would fit within the database and within our classification. Both of these stages use

web pages that we have and also journal searching as whenever a structure is released the expert on that structure is the person who solved it. It then goes into the database and gets released.

My other work deals with databases around Europe that have done other analyses of protein structures or sequences. The idea is to find common ways of linking data between those databases in an understandable way. At the moment it's done in a manual way: you have to go and get the data. The idea is that rather than having your own local copy, you can access people's data in a more machine-controllable way.

B. Source data

B1. Could you confirm the types of electronic source data you produce? (7)

If I write a paper it's about the CATH database - new facilities or entries in the database, or new technologies that make the database accessible in different ways.

Nothing is done specifically to proteins; we classify things. We take information from the databank and enhance it. We're not actually creating anything new, we're classifying.

B2. In what formats are these source data held? (8)

n/a.

B3. How is this data produced?

n/a.

B4. How is this data stored?

CATH is housed on the UCL network and is freely available from anywhere. Anyone can use it and we encourage that. It's a website with downloadable files and there is a degree of software availability as well. CATH has a web page that houses flat files that you can download to a research group and work on that.

Machine controllable methods that I'm doing at the moment is so that people don't have to download something and work on it themselves. We're providing tools so they don't necessarily have to understand how the file is laid out – they can just ask general questions and we can provide the answers.

B5. Generally, how large are the files you generate?

n/a.

B6. You indicated in the questionnaire that you often generate a combination of differing data formats. What formats might be combined? (9)

n/a.

B7. To what extent do you think the data you generate would be useful for other research projects?

CATH is hopefully hugely useful. It's the main point of CATH - to have protein structures that can be broken down into smaller units, so you can get the essence of what the protein structure is built from. It's what researchers in our group and other groups like us rely on.

B8. Would it need any modifying before dissemination i.e. would it be easy for others to use in its raw state?

n/a.

B9. Conversely, why might you wish to access source data generated by other research projects? (10)

We can go so far using the tools we have and we can try and reinvent the wheel, but we recognize that there are other groups out there that have their own particular expertise that they make freely available. So we exploit that as much as possible because it makes our job much easier.

B10. What kind of data?

Usually other people's methods of classification: the PFAM database that classes protein sequences in a different way to us.

MMDB we use as a way of accessing journals.

PDBsum is really useful – he [the site developer] *has taken the PDB structure and other databases such as CATH and summarized them all on one page. It's a great resource to use. Actually he's trying to take PDB structures that have been published and he's trying to write a programme that will access a paper and pull out figures from that paper and put them on that page.*

At the end of the day you can have all these databases adding information about a protein structure but sometimes the best thing to do is just to read the journal article. There is a lot of detail you just can't capture in a database.

The aim of the linking databases together project is to enable biologists with very little knowledge of these databases to ask general questions and to receive detailed information back in a machine-controlled way. We're the middle ground so we have the raw data and we're trying to make meaningful data of that raw data in order to then pass on to people that just want a summary.

B11. How do you find and access it? (11)

See B10.

B12. How might the sharing of this source data be made easier?

This has been the aim from the beginning. Our main competitor is a database called SCOP who are in some ways better than us in what they do as it's more manual. They have a guy who spends a lot of time doing what we get computers to do so he's slightly more accurate in what he does. But I do know a lot of people access our database as we make it more human by concentrating on our webpages, by making navigation easier, and by adding more

functions by being able to pop up different types of pictures and by being able to download an XML version of a page so people can use pages in a machine-readable way.

The biggest thing is advertising to get more people using the site. I've always felt we needed closer contact with the community: at the moment, all our users are anonymous. We would like to be able contact people with changes and also to get more feedback. That two-way communication is missing from CATH.

Biochemistry is slowly becoming a more interactive community, for example I saw something called Biowizard the other day which has PubMed publications and they allow people to write comments against each publication which is unheard of up to now. We do want to do the same sort of thing with our database, so that where we haven't got round to chopping something up and classifying it, we want to make it available so other people can put their own comments in and make suggestions, so that when we come round to doing work manually, we may have someone there with expertise who took the time to make suggestions.

C. Source repositories

Do you submit data to source respositories?

No but we retrieve from pretty much all of them though.

C1. Why did you choose these repositories?

n/a.

C2. How easy was it to submit data?

n/a.

C3. Anything that you particularly liked or disliked about the process?

n/a.

C4. Do you download data from any source repositories?

PDB, Genbank, Uniprot

C5. How frequently?

Every week from PDB, same with Genbank and Uniprot. We have automated scripts.

C6. What were your experiences of this process? E.g. How easy was it to find what you were looking for?

It's very easy. PDB changed formats a couple of years ago but even that turned out to be straightforward as they supported both the original and the new format.

C7. How in your opinion could source repositories be improved?

There isn't a lot really on a general level as they do their job well. There are database-specific things, so for example PDB: the file format that users don't stick to but that's not really their fault as it's the users. There have been some projects like MSD to clean up formats but that's run into problems and we were relying on that to be available but we're still waiting. It's a difficult project.

The biggest pain with all of them is information that changes from week to week. The contents of a structure file will change after its submitted.

Uniprot is supposed to have stable identifiers because that was the problem with Genbank – their identifiers changed although the sequence was the same and this caused problems. There was a collaborative project that said, let's annotate Uniprot with our own expertise and at the end of the project we shall compare annotations and bring them all together into one database. Towards the end of the project they compared data and realized that they had Uniprot identifiers that were the same but different sequences. That should never have happened. When things like that go wrong they go very wrong. You have to make absolutely sure that when you start off you're all using absolutely the same file.

D. Metadata

D1. What metadata fields do you consider important to describe your data? (13)

The classification is the key to the database, in fact the classification IS the database. The full name of CATH is CATHSOLID and each letter represents a level in our hierarchy. Each level has a number and each structure within the database has nine numbers which basically is its classification number. That is our key metatdata. That's CATH.

D2. At what stage are metadata assigned to your research data? (14)

We have the two-stage process of breaking structure up into domain units. Then we go through a classification procedure so that either it happens immediately or it's a slow process of narrowing it down.

D3. Who assigns this metadata? (15)

See D1.

D4. Does your use of metadata vary according to the type of data you submit? i.e. would you use other fields to describe other types of data?

n/a.

D5. Do you know of any standard metadata sets that are used to describe your data?

n/a.

E. Data access and sharing

E1. What measures do you use to make your research data available? (16)

[The database itself i.e. through a source repository.]

E2. What factors would encourage you to share your research data? (17)

[This the raison d'etre of the database.]

E3. What factors would discourage you from sharing your data? (18)

I really doubt it. Any publicity is good. We even collaborate with our competitors - there was a move to find some standards within our classifications to apply to both databases. Anyone can use it.

E4. What kind of formal restrictions do you apply to the release and/or access to your research data? (19)

None.

E5. What actual practical measures or processes do you use to control access to your data? (20)

None. There is no monitoring of users but we do have hit counts.

F. Output repositories

F1. Which output repositories do you use for information to draw on in your research? (21)

Publisher repositories –

Given that people seem to publish in the journals we read anyway, I just go to PubMed. Pretty much all the journals we want have been bought by the university.

With so many journals now you have at pre-print versions and they have the early-view links as soon as the paper has been accepted for publication.

F2. And to source material for teaching you used (22)

n/a.

F3. When making your own research papers available, how do you choose where to publish or deposit?

I'd never make a paper available on my own website as I've always been worried about the legality of it but you often get people emailing in for reprints and then I often send them a pdf. I've been asked by a publisher did I mind them making a paper open access and I said yes

and then I was stuck with a bill. It's a huge amount of money to make your paper freely available.

F4. In which output repositories do you deposit your research publications? (23)

See F3.

F5. Of the output repositories you have used, what were their good or bad points?

I've only ever used PubMed and Web of Science too occasionally for citations. The thing I love on PubMed is the related links.

The PubMed search engine is not that good: you can put in an authors name and some keywords and still not find anything. I suppose if you spent fifteen minutes thinking about it, it might actually be quite powerful but for a first time visitor it's a bit scary.

F6. How could output repositories be improved?

See F5.

F8. Would you consider depositing your research papers in an open access institutional repository?

Yes. I'd have no problem about that.

NB It's something they are actively considering as a lab group.

G. Support

G1. Please would you describe the level of support you receive when using output repositories. This can be from individuals or from online links or advice. (26)

No not used. I actually get a little bit confused with PubMed as it's only one of a number of services on the same page. I've never really explored it.

G2. Do you think you are using output repositories efficiently?

Yes. Actually there is an EBI email list that is forwarded of recent current publications that are all have PubMed links and that have relevance. But that's fundamental. But at the end of the day it's just the journal title and you have to go find everything.

G3. Might there be features of output repositories that you are unfamiliar with?

See G1.

H. Reprise of project aims – Source repositories

H1. If a standard feature of repositories containing source data was the ability to identify and link to the publications developed from this data, how advantageous would you find it? Why is this? (5)

It's difficult for us to say because our source data we get automatically, although we could also automatically pick up papers. We do have references on our website but not papers themselves or even the PubMed IDs for our papers. It makes me wonder if we're missing out on a big chunk of data. There's no reason why we shouldn't provide this service. .

H2. Similarly, if you navigate directly from within an online article or other text to the primary source data from which it was derived how advantageous would you find it? Why is this? (6)

That would be very useful.

H3. Having now considered both source and output repositories, and how they might relate, what functionality do you consider to be missing from the source repositories you have used? (28)

See above.

H4. We are exploring ways of providing links from repositories of source data to repositories of published papers because we believe there is a need amongst researchers to identify published (and pre-published) papers that have made use of their source data. In what way can you identify with this perceived need? (29)

See B12 and E5.

H5. Linking to source data from output repositories will require that an adequate range of metadata is applied to the source data that will persist over time. What sort of difficulties – and solutions to them – might you anticipate when attempting this? (30)

With CATH as we get more and more information a classification may change and that means there will be a potential move within our classification. Versioning in CATH is important but we do have a constant in that each domain has a domain ID and that's about as stable as we can get. When we advise people about linking to our data we always suggest linking to a domain ID.

I. Reprise of project aims – Output repositories

I1. What functionality is missing from the output repositories you have used? (31)

See F6.

I2. We are considering building an interface for output repositories that would let you as a depositor, associate newly deposited publications with the data from which they are derived. In what way might this be of benefit to you or indeed others? (32)

The benefit to you would be that other people were doing it too. You'd definitely want to make it as easy as possible to capture that information.

When we publish papers, as we're providing services rather than a scientific result, often it is a self-advertising anyway. But we're focusing on CATH here: we're also researchers and other researchers also use the database as a whole. Submitting papers is bad enough – there are all these things you have to fill in and you're worrying about it being accepted, so if you were to add more fields I think you would definitely need to make it as painless as possible.

I3. A number of new operations could be supported within an output repository, such as the automatic creation of links, the automatic embedding of source repository data and the presentation of relationships (i.e. showing publications and their source data in adjacent windows). How do you think these features could meet your needs? (33)

One way we would use it in CATH would be at the stage when we were doing the classification. Our source data would be the PDB structure and we're looking for published information about that structure. If we could link from one paper to others that reference that paper, for example if we were looking for functional information or supporting information, then that would be useful.

The other thing that is useful for us is comparing two papers, so that if you have one structure with functional information and you've computed a match between one structure and another, then you may want to see in another paper whether they share the same sort of functional information. That is a project that a lot of people have tried to do before – comparing papers automatically. Even finding two papers that are related is hard enough.

In your given field it's assumed that you should have a grasp on new and historical publications but it's difficult to keep up. That's why I like the related articles in PubMed. If I find a paper that's really interesting, I can just click on the link and I get all the articles related to it. So if there's a new article I've not quite picked up on, then it's there.

If you have CATH as a source and all the papers that are related to CATH, you have a very broad range of papers that are not necessarily related. You'd have to start off by saying, I want to know about a given family and a certain aspect of it and then you want to know what papers. You don't want to know just because the papers are in CATH.

I4. What other features might you expect to be advantageous? (33)

See I3.

J. Reprise of project aims – Potential solutions

J1. A 'dataset knowledgebase' is an online service which allows the creation of two-way links between source and output repositories. It could resolve questions placed in either direction and could also be enhanced through the addition of features such as stored user annotations, quality assessments or ratings and answers to FAQs about specific sets of data held in a repository. What is your opinion of the value of such a concept? (34)

It would be useful if enough people used it.

J2. Are there specific issues you might want it to address? (34)

I've not really used a system where people have done this, so I don't know how well it would or wouldn't work. I would need to see it in action.

One of the things I could envisage being difficult is source data – it's hard to visualize. So often you have a link to the source data but that link may not be helpful if the data itself is

bland. Sometimes it's better to not just link to the source data but to link to other places that have annotated that data e.g. PDB doesn't actually offer a lot, where as PDBsum because it annotates the source data, gives you a lot more information.

And it all depends on the readership. If you read a paper and it's really related to your work, then a lot of the information would be very familiar so any links that were provided wouldn't really be of any relevance. Biologists, say, that aren't as familiar with the things we do may find it very useful because maybe it's a database they've heard of but never used. In terms of CATH and advertising it would be fantastic. CATH is friendly!

J3. Some data repositories are open to all enquirers while others are password protected. If we are expecting to design links that will provide access from open repositories to controlled repositories, we shall need to devise some level of validation and temporary access rights. Could you describe the extent to which this is necessary in the context of your own source data? (35)

n/a.