Molecular Biology Through Discovery

Companion to Sanger & Tuppy (1951)

The amino-acid sequence in the phenylalanyl chain of insulin

1. The identification of lower peptides from partial hydrolysates

Biochem J 49:463-481

I. Introduction

There were several competing ideas in the air during the first half of the 20th century concerning the physical nature of proteins. It was well established that different proteins had different functions. For example, some, like urease, catalyzed specific biochemical reactions. Others, like fibrin, had a purely structural function. It was also well established that proteins were composed of amino acids and that they differed from one another in their amino acid content.

There was a widely held view early on that proteins were not independent molecules but functioned as micelles, aggregates of indeterminate size.[1] This view became untenable as approximate molecular weights of proteins became available in the 1920's and 1930's. Nonetheless, some continued to hold that proteins attained their size through the repetition of constant amino acid units. While fibrous proteins often have a good deal of repetition,[*] it turns out that this is otherwise rare.

Many in the 1930's (see ref. 2) were impressed by reports that the number of amino acids contained by natural proteins were constrained by the equation

$$\text{Number of amino acids} = 2^m \times 3^n$$

Egg albumin, for example, was reported to contain 288 amino acids ($= 2^5 \times 3^2$).[3] As near as I can tell, the actual evidence for this equation is virtually nonexistent, relying on very imprecisely determined ratios amongst amino acid constituents, but the idea took on a life beyond its underlying evidence. In fact, the number of amino acids in proteins is not constrained.

One particularly influential idea, put forth by Dorothy Wrinch[4] and championed by Nobel laureate Irving Langmuir[2] (**Fig. 1**) was the view that that proteins were networks of amino acids, like chained link fences with variable numbers of links. There was no experimental evidence for this idea. Rather, it was motivated by the few structural determinations of proteins that existed, which could be misinterpreted as requiring that the proteins existed in repeating cycles.

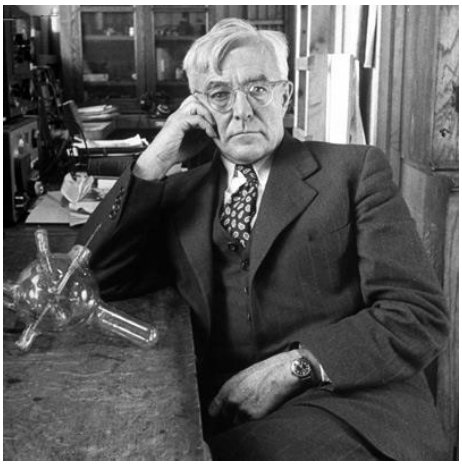

Figure 1: Irving Langmuir (http://www.biography.com/people/irving-langmuir-9373252)

One thing that everyone could agree on was that the field desperately needed hard information regarding the structure of specific proteins. Into this breach stepped Fred Sanger (**Fig. 2)**, who over the course of a half dozen years determined for the first time the linear sequence of amino acids of a protein, insulin.

[*] For example, 94% of the sequence of silk fibroin consists of the repeating unit Gly-X, where Gly is the amino acid glycine and X is some other amino acid (usually alanine) [Zhou C-Z (2001) Proteins 44:119-122]

II. Initial survey of article

With that in mind, find the article by Sanger and Tuppy (if you haven't already), and take a look at it, with the goal of understanding how the first sequence of a protein was determined. Make sure you find the right article (see title of this companion), with the right journal, volume, and page numbers.

You will after cursory inspection that it is a ***research article*** and is structured in a way typical of such articles: **Introduction**, **Methods**, **Results**, **Discussion**, **References**. Most importantly, the article provides observations and the means by which they were obtained, in sufficient detail, or reference to other articles providing the detail, that someone skilled in the art could reproduce the findings.

Courtesy of Dr. F. Sanger, MRC, Cambridge.
Noncommercial, educational use only.

Figure 2: Fred Sanger in the late 1950's.

Introduction

The ideal **Introduction** of a research article accomplishes the following goals:

1. Defines a major problem of interest to the target audience.
2. Through a series of logical steps, presenting prior results along the way, focuses on successively more narrow slices of the major problem.
3. Ends up with a narrow question that can be answered by the experiments that are presented in the body of the article.

Typically, the **Introduction** ends with a simple statement of the narrow question. Sometimes the major finding is also given immediately afterwards, sometimes not. This should not be construed as an arbitrary format but rather the principles of effective communication. If you were involved in a complex project and wanted to explain it to someone far from the field (e.g. your parents or 9-year old child), you might proceed in just this way.

SQ1. Read the Introduction of Sanger & Tuppy (1951) (note that there is no section labeled "Introduction". This section extends from the beginning of the article to the Methods section). Does it address the goals of an ideal Introduction as described above?

SQ2. What is the major problem into which the work of Sanger & Tuppy fits?

SQ3. What is the narrow question that is the focus of the article?

After going through the **Introduction**, you might wrongly conclude that you're too stupid to read this article. A better way to put it is that there is a mismatch between the authors target audience and us, or the article is not well written. I think both are true. Fred Sanger is the only biologist ever to win two Nobel prizes, so there's no question of his fire power. Nonetheless many excellent researchers are not also excellent writers, and unless you're prepared to throw away excellent research, you're going to have to live with that.

The fact is, the **Introduction** does not present a major problem of any sort. Rather it jumps right into the minutiae of methodology, of great interest to those in the field at the time but distinctly less interest to those of us reading the article 60 years later. That is the reason I wrote my own introduction (previous page). You might also look at Antony Stretton's review of the article.[5]

Despite the deficiencies in scope, the article is remarkable in the clarity with which it presents the observations that led Sanger & Tuppy to deduce much of the amino acid sequence of insulin. I advise you go into it with the aim of figuring out how someone deduced that sequence, to the extent that you can do it yourself. Don't worry about the rest.

Methods

As usual, skip this section. Maybe the time will come when you find that there's something in it that you need. Maybe not.

Results

I would ordinarily expect (or at least hope) that this section begins by repeating the narrow question of the article and presenting the overall strategy used to address the question and the principle behind the method employed by the first step of the strategy.

SQ4. Read the beginning of the Results section. What is the overall strategy the authors used to determine the structure of Fraction B of insulin?

SQ5. And what is Fraction B anyway?

Well, newbies are not going to get a lot of help from this article. We're going to have to take a step back before we get into the results.

III. Experimental strategy used by Sanger & Tuppy

Let me step in and give it a try. First of all, this article focuses on just the larger of the two unique peptide chains of ox insulin, called (equivalently) Fraction B and phenylalanine chain (after its first amino acid. Insulin has two copies of each of the two chains (**Fig. 3**). I must stress that Sanger & Tuppy did not know the structure shown, nor even that insulin had a unique sequence. These insights remained for them to discover.

The strategy employed by Sanger & Tuppy to determine the sequence of the phenylalanyl chain of insulin proceeded in the following steps (**Fig. 4**):

1. Step 1: Isolate the phenylalanine chain of insulin (also called Fraction B). In brief, the cysteine-cysteine bonds are broken (using the oxidant performic acid), converting the cysteines to cysteic acid (S&T write it $CySO_3H$). Then the more basic B chain is separated from the more acidic A chain.

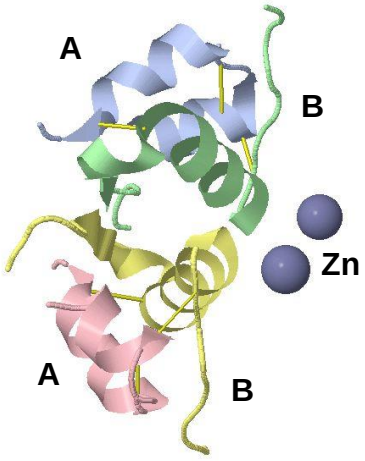

Figure 3: Structure of bovine insulin (PDB 4e7t). The two glycine chains (Fraction **A**) are shown in blue and pink. The two phenylalanyl chains (Fraction **B**) are shown in green and yellow. Zinc, a natural component of insulin is shown as purple spheres. Covalent bonds between cysteines are shown as thin yellow rods.

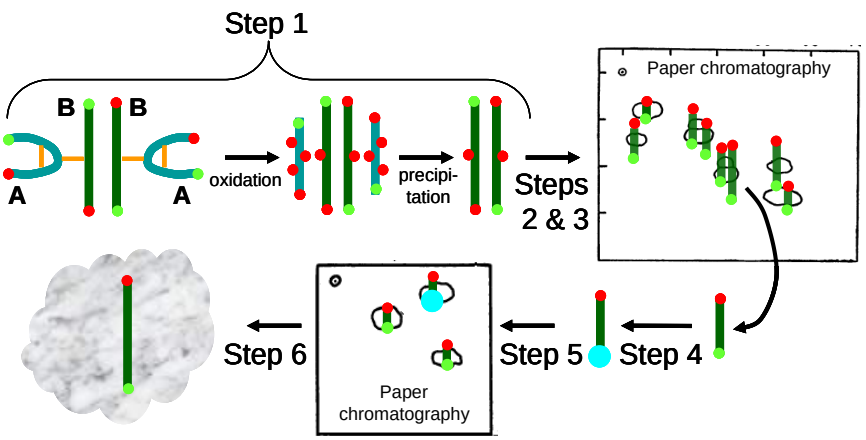

Figure 4: Strategy to deduce structure of phenylalanine chain of insulin. The steps shown are described in the text. The thick cyano and green lines represent the glycine (Fraction **A**) and phenylalanine (Fraction **B**) chains, respectively. Orange bars are Cys-Cys bonds. Red circles are acidic groups. Green circles are amines. The large cyan circle represents dinitrophenol used to derivitize the N-termini.

2. Step 2: Fragment the B chain at random into small peptides. In this article, fragmentation was done by cooking the protein in strong acid or strong base. It is important to understand that the procedures did not completely break the protein down to its constituent but rather nicked the protein chain at random locations.
3. Step 3: Separate the small peptides so that each one can be individually analyzed.
4. Step 4: Label the N-terminal amino acid of each small peptide.
5. Step 5: Identify the amino acids of each small peptide, by breaking up the peptide into individual amino acids. Separately, do the same with unlabeled peptide (skipping Step 4). Determine N-terminus of peptide by looking for spot missing in labeled sample compared to unlabeled sample.
6. Step 6: Piece together all of the peptides to form the complete sequence of Fraction B.

SQ6. Try SQ4 and SQ5 again.

Every experiment has its limitations, and this one is no exception. The fragmentation step at high temperature (Step 2) destroys some amino acids, converting glutamine to glutamate and asparagine to aspartate. It is therefore impossible with this method to distinguish the two. Paper chromatography used to separate peptides (Step 3) and amino acids (Step 4) is sometimes incapable of fully separating all elements in the sample.

Paper chromatography is used in Step 3 to separate peptides from one another and again in Step 5 to separate amino acids. It's the technique on which all the results in this article relies, so it is incumbent upon us to grapple with it. Now is a good time to look at the presentation *Paper Chromatography*, available from the calendar and from the **Protein** unit web site.

SQ7. Consider Fig. 4a in Sanger & Tuppy. What are those numbered white blobs?

SQ8. In great molecular detail, why do spots #9 and #13 have about the same x coordinates but different y coordinates?

SQ9. Are there amino acids found in insulin that are absent in the phenylalanine chain of insulin?

IV. Results in Sanger & Tuppy

IV.A. Overall consideration of the results

Now you're ready to go through the **Results**. Note that I'm not suggesting that you actually ***read*** the **Results** section, just go through it, noting what there is so as to judge what part of it is worth reading. "Worth reading" means "important in helping you achieve your goal of finding out how the first protein sequence was determined." To do that, make a brief outline of what results there are.

SQ10. Write an outline of the results section.

This should have been real easy, so long as you don't feel compelled to understand what was written. The outline could have been a few lines:

> For each fraction from B1α through B5γ
> For each table from 3 to 13
> For each figure from 5 to 16
> Sanger and Tuppy present in a table and a figure the analysis of peptides present in a fraction.

SQ11. But what are those experiments? Make a list of each fraction symbol (e.g. B1α) and write next to it what how that fraction differs from the others.

How do you find out what these fractions are? If you find nothing in the **Results** section or in the figure legends or table footnotes, then you need to grit your teeth and see if the **Methods** section has anything to offer. By the end of SQ11, you should have a useful table, so useful, in fact, that I wonder why Sanger and Tuppy didn't provide it themselves.

Since most of the tables from 3 to 13 appear to be about the same in format, and the same for all the figures from 5 to 16, all we need to do is to figure out one set. We might as well choose the first: Fraction B1α, Table 3, Figure 5.

SQ12. There are eight numbered blobs in Fig. 5. What do they represent? How are they distinguished from the 13 numbered blobs in Fig. 6?

SQ13. What is the relationship between Fig. 5 and Table 3?

SQ14. What is the relationship in Table 3 between the column labeled 'Deamination and hydrolysis' and the column labeled 'DNP treatment and hydrolysis'? (You can answer this even if you have no idea what these labels mean) (However, if you have no idea, your understanding of the results would be greatly enhanced if you found out – see below).

SQ15. What is the relationship between the content of the three columns under the heading 'Strength of amino acid after' and the column labeled 'Structure'? For example, justify why in Table 3, spot number 4, the structure is given as Val.CySO$_3$H.

SQ16. Can the structure given in Table 3, spot number 4 be reconciled with the structure given in the same table, spot number 6?

That's pretty much what the article has to say about insulin (except for the process of putting all the information together!).

IV.B. Determination of amino acid content of a peptide and amino acid order

All the chromatographs in this article show spots consisting of peptides, the outcome of Steps 2 and 3 of **Fig. 4**. Yet, tables associated with each chromatograph list the amino acid content ("Amino-acids present") and the amino acid order ("Structure") for each spot. Where does those conclusions come from?

The amino acid content of a spot's peptide was determined by eluting each spot and hydrolyzing the peptide to individual amino acids. The collection of amino acids were chromatographed (Step 5 in **Fig. 4**). Comparing the position of the spots to those of Sanger & Tuppy's Fig. 4 permitted identification of each amino acid.

The amino acid order of a spot's peptide was partially determined by reacting the eluted peptide with dinitrophenylfluoride (DNPF), which reacts with free amino groups, notably with the amino group at the N-terminus of a peptide. Hydrolysis of the derivitized peptide yields the same amino acids except for the N-terminal amino acid. Comparing the chromatograms of both hydrolyzed samples indicates the N-terminal amino acid as the amino acid whose spot moves.

If peptide has more than two amino acids, then this method can determine the position of only the first. The positions of the remaining amino acids remains ambiguous, represented by Sanger & Tuppy as sets of amino acids (e.g. [Gly,Ala]) to indicate unknown order.

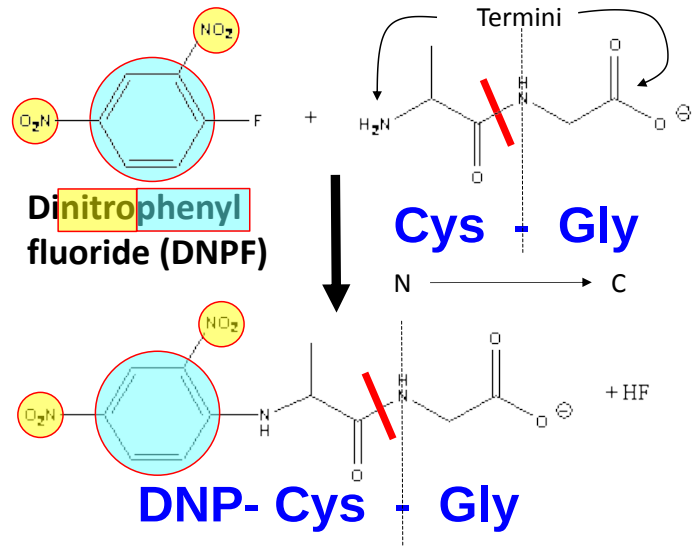

Figure 5: Determination of N-terminus of a peptide. The N-terminus of a peptide is the side that has the free amino end. The C-terminus has the free carboxyl end. Hydrolysis of the Cys-Gly peptide produces Cys + Gly. Reacting the peptide with DNPF modifies the N-terminus, so that hydrolysis produces the DNP-Cys (much bulkier than Cys) + Gly.

SQ17. Draw the two chromatograms you imagine that led to the conclusions in Table 3, Spot 1.

SQ18. The N-terminus of Table 3, Spot 8 is more reliably determined than that of Table 3, Spot 5. Give two reasons why.

There's only one more set of pertinent information, which does not come from the experiments reported in this article but is provided in the article nonetheless. That's Table 14.

SQ17. How might the information in Table 14 be useful in putting together the sequence of insulin, Fraction B, from the results reported in Sanger and Tuppy? Specifically, how might the first line of that table (concerning cysteic acid) help in interpreting the results given in Table 3?

SQ18. What are the limitations on the data shown in Table 14?

V. Parting thoughts

- Sanger and colleagues continued to elucidate the amino acid sequence of insulin, completing the A chain in 1953. By the end, they had found no evidence of branched chains or repeating patterns of amino acid. They In an article describing their work,[6] they speculated that proteins in general have no such patterns. Rather any sequence of amino acids is possible, and that sequence determines the structure and function of the protein. This view, largely accurate, was enormously influential on subsequent researchers.
- It would be nice if research articles explained themselves in a language that the world can understand, but this is not often or indeed usually the case.
- In many, possibly most research articles, it is easier to understand the experiment and the results than to understand the words of the authors who are trying to explain them. I generally head straight to the tables and figures.
- Scientific progress depends on the combination of various talents – the breathtaking conceptual leaps of someone like Francis Crick would not get us very far without technical brilliance of someone like Fred Sanger.
- Paper chromatography is no longer used to analyze peptides, but the concepts that underly the technique can be found in a great number of biochemical procedures that will continue to be used well into the future.
- I did not lead you in these notes through the process of Step 6, piecing together the peptides to arrive at a (nearly) complete sequence of insulin, Fraction B. We'll go through that process as a group in class.

References

1. Dickerson RE (2005). *Present at the Flood: How Structural Molecular Biology Came About.* Sinauer Assoc, Sunderland MA. p.4
2. Langmuir I (1939). The structure of proteins. Proc Phys Sci 51:592-612.
3. Bergmann M, Niemann C (1937). On the structure proteins: Cattle hemoglobin, egg albumin, cattle fibrin, and gelatin. J Biol Chem 118:301-314.
4. Wrinch DM (1937). On the pattern of proteins. Proc Royal Soc London A 160:59-86.
5. Stretton AOW (2002). The first sequence: Fred Sanger and insulin. Genetics 162:52327-532.
6. Sanger F, Thompson EOP (1953). The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J 53:366-374.