

Using probability: A cautionary tale concerning BRCA1

It is desirable to make database searching as much of a “pushbutton exercise” as possible, without requiring great expertise from the user. But effective database searching often still requires a judicious mix of biological and statistical insights. Recent experience with the BRCA1 gene illustrates this point.

Familial breast and ovarian cancer has been linked to mutations in two genes, BRCA1 and BRCA2, which have both been precisely located, cloned and sequenced. Finding and cloning these genes was a major milestone in cancer research, but therapies or preventions for these cancers require an understanding of the proteins that these genes code for. In the case of BRCA1, there was some early hope that such an understanding had begun, after a database search found a PROSITE motif (see Section ??) in the amino acid sequence derived from the BRCA1 gene [2, 5]. A substring of the BRCA1 sequence matched the consensus sequence for a family of proteins called *granins* (its consensus sequence is displayed in Section ??). Granins had not previously been connected to breast cancer, but based on the database match, several experiments were done that supported a role for granins in breast cancer. However, other experiments contradicted those conclusions. What was exciting and caused a great deal of attention, is that granins are proteins that are secreted outside of the cells where they are made. This is in contrast to most known cancer-related proteins, which work inside the cell (or inside the nucleus of the cell), where it is hard to get at them or to supplement them. If the BRCA1 protein does in fact function like the known granins and is secreted outside the cell, successful strategies against oncogenic mutants of the BRCA1 protein seemed much more likely.

Unfortunately, a deeper examination of the original PROSITE search concluded that “the granin motif has no statistical support” [3]. The first paper [2] had reported that “the probability of the granin consensus occurring by chance in BRCA1 is approximately ... 0.00175”, and hence concluded that the granin match reflected an important biological phenomenon. However, without prior evidence that granins are involved in breast cancer, the computed number (which is a reasonable approximation to the probability that the granin motif occurs by chance in the BRCA1 sequence) does not answer the right question. The more informative question is: what is the probability that *some* PROSITE motif (as complex as the granin motif) would occur by chance in BRCA1. That probability was reported in [3] to be 0.87. This high number is due to the fact that PROSITE contains well over one thousand motifs, that the granin motif is only ten positions long and three positions are wildcards, and BRCA1 is a relatively long protein (1863 residues). In such cases, additional work is needed to make an effective use of the sequence and motif databases. For example, one should look for more extensive similarities between the BRCA1 protein sequence and the sequence containing the matched motif, beyond the motif region itself. Or one should first identify short stretches of the BRCA1 protein sequence where a matched motif would more likely reflect a real biological phenomenon, and where probability computations would be more informative.

The above approach was followed in [3]. They first identified six regions of the BRCA1 sequence with a predicted globular structure, since globular regions are more likely to contain conserved functional motifs. Matching each of these six regions against a sequence database, they found a moderate match between a 202 residue long globular region and an analogously located region (at the C-end of the proteins) in a human protein, 53BP1, known to bind to the universal tumor suppressor p53. Moreover, the match consisted of two distinct segments that were separated by almost the same number of residues in each protein (100 and 97). The similar position and shape of the matching substrings suggest that the match has some biological importance. The use of such evidence is sometimes important in augmenting pure probability computations. The biological plausibility of a connection between BRCA1 and p53 (which is associated with about half of all cancers) is strengthened by the fact that cancer causing mutations in the BRCA1 gene more frequently occur in the 3' end of the gene, which corresponds to the C-end of the BRCA1 protein. This line of reasoning, and additional evidence presented in [3] lead the authors to suggest that the BRCA1 gene is associated with a well-established cancer agent and not a granin¹.

0.1 Importance of searching protein with protein

Overwhelmingly today, new protein sequences are obtained by sequencing the underlying DNA in the gene (or the mRNA) that codes for the protein. Therefore, most of the entries in the “protein” databases are actually “derived amino acid sequences”, and their originating DNA sequences are contained in the DNA databases. So, when a searcher wants to search for a similar protein sequence, the searcher can often use either the DNA sequence of the query protein to search the DNA databases, or can use the translated sequence to search the protein databases. There are many reasons to maintain and explore the underlying DNA sequence, but to detect similarities in highly diverged sequences, the standard advice (from Doolittle [1] in this case) is to

Translate those DNA Sequences!

Some beginning sequence comparers are under the impression there is more to be gained by searching the actual DNA sequence rather than the amino acid sequence derived from it. Such a course is greatly mistaken ...

The reason to translate is that more sensitive and informative comparisons are possible between amino acid sequences than are possible between DNA sequences, *using the common alignment methods and objective functions* of the type detailed in this book². All (and more) of the “information” contained in a derived amino acid

¹They did offer one surprising suggestion. Searching the database with a profile constructed from the most conserved parts of BRCA1 (from human and mouse) and 53BP1, they found a protein in yeast that might be structurally related to p53.

²The reason to translate is often stated as something like “amino acid similarities are much better preserved through evolution than similarities in the underlying DNA”. In fact, that is the statement

sequence is contained in its underlying DNA sequence. But that DNA information is not in a form that directly allows common alignment methods and objective functions to be the most effective.

Derived amino acid strings allow more meaningful alignment by using scoring matrices that reflect the evolutionary, biological or chemical similarities of specific amino acid pairs. For example, replacement of leucine by isoleucine is generally considered a minor change and that fact is reflected in the PAM and BLOSUM matrices. But alignment scores derived by summing the scores from matches and mismatches of the underlying nucleotide pairs would not capture the close relationship of leucine and isoleucine. Similarly, most amino acids are coded for by more than one codon, and two codons differing only in their third position often code for the same amino acid. This kind of correlated, third-position effect would be hard to correctly model in alignments of DNA using alignment methods discussed in this book. The mismatch score of a pair of nucleotides must reflect the appearance of those nucleotides in many different amino acids. And since there are only sixteen different nucleotide pairs, the sum of pair scores over a codon would rarely have the specificity of an amino acid substitution score from a matrix for 400 pairs. There are, however, some efforts to get the best of both representations, by using derived amino sequences, but remembering the specific codons in the underlying DNA. This ideally requires a codon substitution matrix of size 64 by 64.

References

- [1] R. F. Doolittle. *Of Urfs and Orfs: A primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA., 1986.
- [2] R. Jenson, M.C. King, and J. Holt et al. Brca1 is secreted and exhibits properties of a granin. *Nature Genetics*, 12:303–308, 1996.
- [3] E. V. Koonin, S. F. Altschul, and P. Bork. Brca1 protein products: Functional motifs. *Nature Genetics*, 13:266–268, 1996.
- [4] M. Max, P. McKinnon, K. Seidenman, R. Barret, M. Applebury, J. Takahashi, and R. Margolskee. Pineal opsin: A nonvisual opsin expressed in chick pineal. *Science*, 267:1502–1506, 1995.
- [5] P. Steeg. Granin expectations in breast cancer? *Nature Genetics*, 12:223–225, 1996.

I had in a draft of this book before Stephen Altschul pointed out that the critical issue is not what information is conserved in the sequences, but how well common alignment and statistical methods can use that form of information. And, in fact there are cases where the DNA sequences are claimed to be more highly conserved than the translated amino acid sequences[4].