



Protein Potential

A new type of relationship is bringing fresh potential to bioengineering, combining artificial intelligence and synthetic biology to give promising protein pharmaceuticals

By Dr Claes Gustafsson at ATUM

Structure-activity relationship (SAR) studies have been instrumental in bringing small molecule drugs to the market over the past few decades. However, as protein pharmaceuticals replace small molecules as the primary modality for drugs, SAR as we know it is being usurped by the new, technology-driven sequence-activity relationship. This has several unique properties that make it ideal for machine learning approaches to engineering new protein pharmaceuticals, and the time to exploit this bioengineering pathway is now.

SAR

SAR and its mathematical sibling, quantitative SAR (QSAR), have long been used to describe the relationship between the chemical properties of a small molecule

and its corresponding functional activity. It is based on the premise that, given a sufficiently large dataset of related molecules and their quantitated functional output (eg binding to a target protein), we can build QSAR models that predict the functional output of similar molecules not present in the dataset. SAR is often also used to predict new molecules that are improved over the molecules in the training set and can also be used for classification applications.

First established in the 1960s by Professor Corwin Hansch at Pomona College in California and his collaborators, the QSAR concept has since become the gold standard. Today, QSAR is taught in undergraduate level chemistry classes and commonly used in drug discovery and toxicity prediction. The success and

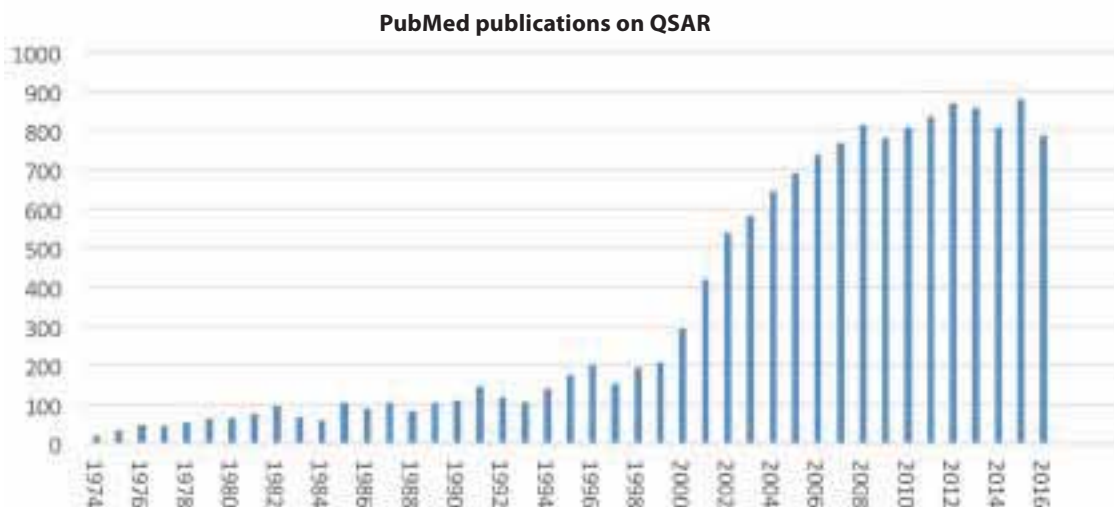
influence of QSAR is evident from the number of publications on QSAR in PubMed – approximately 800 per year since about 2008 (see Figure 1). The majority of these publications are focused on applying QSAR to small molecule drug discovery.

Biotechnology is revolutionising current drug discovery. Small molecules are no longer where the drug industry is investing its time and money. Instead, over the past few years we have seen a tidal wave of protein pharmaceuticals arrive at the marketplace. In 2016, 7 of the top 10 best-selling drugs were protein

Keywords



Protein pharmaceuticals
Protein sequences
Sequence-activity relationship
Structure-activity relationship
Synthetic biology

Figure 1:
Number of publications in PubMed discussing QSAR



pharmaceuticals, and sales of biologics reached a record high of \$163 billion, with nearly two-thirds of all biologics sales attributable to antibodies. Just six years earlier, not a single one of the top 10 drugs was a protein pharmaceutical (see Table 1). Proteins in the form of natural antibodies, enzymes and engineered new formats are expected to be the primary modality of therapeutic drugs for the foreseeable future.

Also changing is the methodology; no longer is QSAR a design tool for protein pharmaceuticals. Protein engineering has instead been approached from two diametrically opposed directions: rational design and directed evolution. The former attempts to understand protein structure

 Directed evolution attempts to find a desired solution by testing a large number of semi-randomly generated variants, typically using various evolutionary-based algorithms 

and function from a mechanistic level so that any desired change can be affected by calculations from first principles. This is usually accomplished by building a crystal structure of the protein drug bound to the ligand (a non-trivial task), followed by rationally designed mutations and the testing of these until the required molecular behaviour is achieved. Instead, directed evolution attempts to find a desired solution by testing a large number of semi-randomly generated variants, typically using various evolutionary-based algorithms. Because both rational design

and directed evolution (in their many alternative formats) have shortcomings and advantages, both are now often used in conjunction for improved success rate.

Protein Pharmaceuticals

Instead of using small molecules' quantitative structure parameters as drug-derived variables, protein pharmaceuticals allow for a defined qualitative measure in using the primary amino acid sequence itself as variable. Even though the available protein sequence space is, in principle, infinite, every residue in a protein is limited to only 1 of 21 available qualitative options (the 20 natural amino acids and the lack of an amino acid). Unlike the typical quantitative descriptors in small molecule QSAR (eg size, charge, polarity and so on) where infinite alternatives are available, protein sequences deal with a predefined set of qualitative descriptors. From a sequence-space exploration standpoint, the 21 options at any residue provide a finite number of possibilities to explore – not counting post-translational modifications. Arguably, that finite space is large. With 21 options at each residue in

Table 1:
Top 10 selling drugs 2010 and 2016. Protein pharmaceuticals are marked in yellow

	2010	2016
1	Nexium	Humira
2	Lipitor	Harvoni
3	Plavix	Enbrel
4	Advair Diskus	Remicade
5	OxyContin	Rituxan
6	Abilify	Revlimid
7	Singulair	Avastin
8	Seroquel	Herceptin
9	Crestor	Lantus
10	Cymbalta	Prevnar

a protein, the space is 21^N , where N is the number of residues in the protein. Despite the large possible protein sequence space, almost all of this is non-functional for any given property, in most cases failing even to fold into a defined structure. Naturally occurring proteins that are folded and have fitness for one or more functional property occupy extremely small regions of this space, much like tiny islands within the total available Pacific Ocean of sequence space.

Building Better Biology

So how do we find the islands of good biological activity? Fortunately, Genbank and other public databases provide very large datasets of naturally occurring sequences. As of June 2017, Genbank consists of approximately 250×10^9 base pairs, and whole genome shotgun submission (WGS) is almost an order of magnitude larger. Both of these are growing rapidly. The information provided by the genomic databases can be used as beacons to outline the contours of the small functional islands in the sequence space ocean. Similarly to how Google uses deep learning artificial intelligence (AI) to recognise images of cats by showing the AI millions of cat pictures, we can now show AI millions of gene/protein sequences derived from Genbank/WGS and use the resulting algorithm to tell us which non-natural sequences are likely to be correctly folded and look like a 'real' protein. By limiting the search for improved protein sequences to islands of 'real' proteins, we can drastically improve the success rate of bioengineering.

The precise and qualitative nature of amino acid sequences, all the big data present in public

databases and the ability to make any biological sequence using synthetic biology makes protein engineers uniquely positioned to apply machine learning. As early as 1993, Svante Wold and collaborators at Umeå University used partial least squares projections (PLS) to latent structures and synthetic 68bp DNA fragments to build a quantitative sequence-activity model to predict the SAR of *Escherichia coli* transcriptional promoters. The model was validated by two DNA fragments, which were predicted to be more potent promoters than any upon which the model was based. The optimised structures were experimentally verified to be strong promoters *in vivo*.

A few years later, scientists from Freie Universität Berlin applied a set of 90 related peptides as a training set to a neural network. The resulting algorithm correctly predicted whether new peptides would be active or not. Following on the nucleotide and peptide work, it was easy to see that proteins were next. Aita and Husimi from Saitama University in Japan published a series of papers in the late 1990s and early 2000s that laid the foundation for sequence-activity relationship space exploration in proteins. The Japan team built SAR maps of dihydrofolate reductase, prolyl endopeptidase and several other enzymes. Consensus was building – the adoption of machine learning-driven navigational tools was going to transform the SAR of small molecules to the sequence-activity relationship for protein engineering, but not quite yet. The cost and complexity of making synthetic genes, the limited size of big data and the lack of computational infrastructure for AI/machine learning would not yet make sequence-activity

relationship for protein engineering the new gold standard. Today, that has all changed.

We are now living in exciting times, when the first generation protein pharmaceuticals (primarily regular antibodies that did not require much engineering) are rapidly being replaced by non-natural bifunctional antibodies, antibodies conjugated to small molecule drugs (ADC), those fused to cytokines, single chain antibodies and many other therapeutic protein formats that do not exist in nature. These new formats show incredible therapeutic potential, but are often limited by poor protein expression, instability, aggregation and many other developability features that are hard – if not impossible – to address by directed evolution or rational design. We will need all of the tools in our bioengineering tool chest to move these molecules from promising research papers to robust commercial drugs. The time for sequence-activity relationship is now.



As ATUM's Co-Founder and Chief Commercial Officer, Dr Claes Gustafsson oversees most of the company's external communications. Prior to co-founding ATUM, Claes was Scientist and, later, Manager

at Maxygen Inc where he led, managed and collaborated with key strategic teams for more than five years. He also held a scientist position at Kosan Biosciences, as well as a number of research, teaching and post-doctoral positions at UCs Santa Cruz and San Francisco, US, and at University of Umeå, Sweden. Claes holds 43 issued US patents and has published over 40 scientific papers. He received his PhD in molecular biology/biochemistry from the University of Umeå, Sweden, where he studied translation under Professor Glenn Björk.

Email: cgustafsson@atum.bio