Typical representations for proteins depict them as static structures with atoms in fixed locations. However, in reality proteins are highly mobile and flexible. With Dr Ray's MoFlow approach, the lab has attempted to use techniques similar to time lapse photography to show how a protein moves.

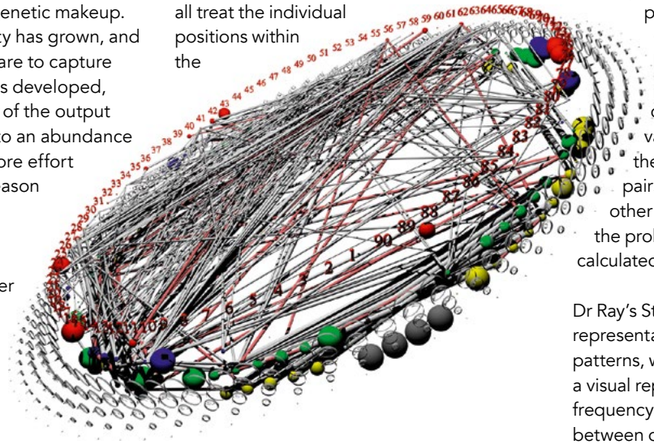
Exploring proteins through visualisation

Proteins are the building blocks of cells. Understanding their fundamental biology and the gene sequences that code for them is of crucial importance for advancing biotechnology and medicine. However, the standard approach for representing proteins and their components has remained two-dimensional throughout the years, causing most analysis software to ignore the importance of interactions between different regions of the protein sequence. **Dr William C Ray** at The Ohio State University has been working to change this though, by developing new statistical analysis and visual representation software, called MAVL and StickWRLD.

One of the fundamental tasks in bioinformatics is the examination of nucleic acid and protein bio-sequences. Researchers use software to analyse and display related patterns within groups of sequences, for example through sequence alignments that highlight the degree of similarity in genetic makeup. However, as data complexity has grown, and the power of analysis software to capture details within sequences has developed, the intuitive interpretability of the output has decreased. This is due to an abundance of information, requiring more effort to decipher. Also, for this reason some of the more powerful results have been primarily designed to be used for further digital analysis, rather than communicating the patterns in the data clearly to a human.

NEGLECTED BIOLOGICAL COMPLEXITY

Despite powerful bioinformatics tools for detecting similarities between nucleic acid or protein sequences having been available for some time, software has been insufficient for meeting all researchers' needs. Most analysis tools suffer from the same fundamental statistical shortcoming: they all treat the individual positions within the sequence as separate in identity from every other position. This is a major problem for biological analysis, as, in reality, the picture within protein sequences is far more complex than that. Conventional software generally doesn't give the user any information about how regions of the sequence may have inter-positional dependencies to one another. From a statistical point of view, these dependencies mean that not all positions in the dataset obey the same strict consensus rules that are usually applied to each data-point.



Dr Ray's MAVL and StickWRLD systems produce a composite representation of data presented to the user as a three-dimensional, interactive model

To resolve this issue, Dr Ray and his team at The Ohio State University have developed the MAVL (Multiple Alignment Variation Linker) and StickWRLD, which produce a composite representation of data presented to the user as an interactive, three-dimensional model. This new software utilises a recently developed statistical technique known as the Conditional Random Field (CRF), which mimics the process of evolution selecting for cooperating components in proteins. Therefore, for example, the descriptions of proteins produced using the CRF can be used to predict whether a mutation in a protein would render it non-functional.

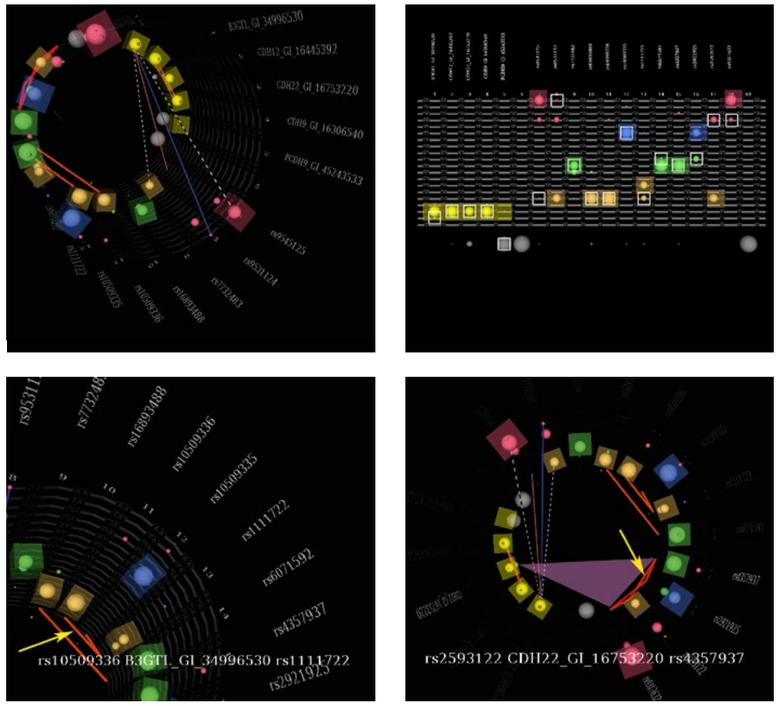
INTRICATE AND INTERPRETABLE ANALYSIS

MAVL calculates values for the association between each possible shared pair of sequence positions and identities. The program compares the value that would be expected between the bases – if their identity in each position is defined by the overall consensus – with the calculated value for each. Deviations from the consensus reveal positions of pairs where their influence on each other is a factor in their identity, with the probability of their presence being calculated to relay potential significance.

Dr Ray's StickWRLD then produces representations of aligned sequence patterns, which simultaneously provide a visual representation of both positional frequency characteristics and the interactions between different sequence positions. The visual models of the networks are determined from the algorithm results finding functionally related sequences based on evidence that a pair of amino acids or nucleotides have a relationship with each other, where if one position changes the other must also be altered.

Where the software identifies a relationship between two positions that cannot be explained by the general consensus, it presents the connection with a line. Other information is contained within this line's characteristics, with the thickness indicating the strength of the relationship and colours indicating its nature.

As well as creating software capable of accurate DNA and protein sequence comparison, Dr Ray and his team wanted to provide scientists with a user-friendly, intuitively interpretable visual representation of the data output. Therefore, they developed StickWRLD to be fully interactive, building an interface that allows the user to interact with the model and explore it from different perspectives. The user can also configure the parameters as they see fit, to control the stringency of the calculated statistics.



StickWRLD has evolved from being able to represent how different protein residues relate to each other, to representing, for example, how different biomarkers, genetic sequences and environmental factors impact clinical outcomes. In these figures, the four columns with yellow spheres correspond to variations on clinical outcomes, and the other columns correspond to possibly correlated or causative clinical variables.

DISCOVERING CLINICAL APPLICATIONS

Since developing StickWRLD, Dr Ray and his team have realised that their algorithms for identifying networks of interrelated and changing features could be applied to much more than evolutionary analysis of nucleic acid and protein sequences. They have taken the idea behind identifying and visualising networks of co-evolving residues in molecules, and applied the concept to clinical data.

To test the system for this application, the researchers used a dataset constructed from several human sets of genetic quantitative trait locus (eQTL) data, to see if they could efficiently search for significant mutations within it. Cancer drug development typically employs high-throughput screening of cancer cell lines, which has yielded an abundance of data on the genetic sequences of a range of cancer cell types. These have been analysed for genomic characteristics that can be related to an array of factors, including drug responsiveness. This has allowed researchers to begin correlating genetic identities with not only phenotypes but also changes that are induced by pharmaceutical intervention. However, this work has been hindered by the fact that as the datasets grow larger their analysis becomes increasingly complex and unmanageable.

PERSONALISED MEDICINE

Through their experiments, Dr Ray and his team have demonstrated that their software has the capacity to overcome this problem and provide researchers with more intuitively accessible models of the statistical characteristics of their data. In doing so, they are moving us towards personalised medicine, by providing an easier method to predict the response a compound may elicit based on a cell's individual genetics. This also helps to streamline the drug development process. Predicting which drugs likely will or won't elicit a response in a particular cancer cell line can help refine the list of candidate chemotherapeutic compounds prior to screening for response *in vitro*.

Dr Ray and his colleagues have also applied StickWRLD to data with a focus on facilitating personalised precision medicine. If medicine is to tailor treatment to individuals, rather than assume what works best for an individual patient based on the majority, clinicians will require tools that can examine how this person compares to, and may differ from, a population. These types of tools are only just beginning to come into existence and Dr Ray is leading the way in demonstrating their potential. His StickWRLD method interactively visualises and compares thousands of sets of patient data, into which an individual's data can be projected for analysis.

StickWRLD, including MAVL, is freely available as an online tool for the research community, available at www.stickwrld.org.

If medicine is to tailor treatment to individuals, clinicians will require tools that can examine how this person compares to, and may differ from, a population

Q&A

Why did you decide to embark on the StickWRLD project?

StickWRLD was born back in 1996, when I was a graduate student. My advisor kept asking why BLAST produced incorrect results for some searches. The answer was invariably interdependencies amongst sequence positions, due to DNA/RNA structural features. Of course, sequencing was in its infancy back then, so we didn't know that. I got frustrated analysing the searches by hand to detect the interdependencies, and wrote the first version of StickWRLD so that my advisor could analyse the sequences himself. So, in short, what has turned out to be my most enduring research theme, owes its beginning to an effort to get my advisor to quit bugging me.

What would you say are the most difficult aspects in developing this sort of statistical software from scratch?

By far, the hardest part is recognising the real question, in a formally definable way. Most of the analysis software that's out there, answers a question that's "kind of like" what the researcher needs to know, rather than the actual question. Sometimes this is because of misunderstanding, but usually it's because the question isn't well defined. I was extremely lucky with StickWRLD. I knew exactly what I wanted the algorithm to show, so it was reasonably easy to code up. Despite this, I took several years to realise that what StickWRLD calculated was well-defined statistically, rather than simply an ad-hoc solution.

Have you received much interest in StickWRLD from the medical community thus far?

It's hard to see a grimace in text... We have a handful of clinical collaborators who are making good use of StickWRLD. We've experienced a depressingly consistent pattern with many others: we create a preliminary analysis of their data and hear "We already knew that, and that, and these things over there. This stuff here we don't know about – how do you know that's true?", and then we never see them again. I think we're engaging them at the wrong time. They're coming to us trying to use StickWRLD as a validation tool, when it's really much better at hypothesis generation. The kicker is that the things StickWRLD gets right in five minutes, that they already knew, cost them years of work to find originally. We need to start engaging clinicians earlier.

Are there any other applications that you think could benefit from the use of StickWRLD?

StickWRLD is useful almost anywhere that has high-dimensional data in which it's interesting to explore patterns in how the variables are correlated. One of my former graduate students loaded up a huge archive of American football play-by-play data and game outcome statistics, so that he could optimise fantasy-football picks. A colleague is currently using StickWRLD to identify cultural drivers of bad business-management practices. Really, anywhere you find that "thing A affects thing B, with a conditional effect that depends on the value of thing A and thing B", StickWRLD is a better analysis approach than more traditional methods like Mutual Information.

What are your hopes for the future of your StickWRLD project?

From an application standpoint, we'd love to identify more domain applications and develop new collaborations that help us expand StickWRLD's utility. From an algorithm standpoint, we're very interested in automating additional varieties of statistical calculations on top of StickWRLD visualisations. In particular, Probabilistic Graphical Models are very popular right now in the Machine Learning and Big Data communities, and StickWRLD is a surprisingly good visual editor for many types of PGMs. We'd love to expand that utility by embedding things like PGM parameter annealing directly into StickWRLD.

Detail

RESEARCH OBJECTIVES

Dr Ray's research focuses on understanding the disconnect between complex biological systems and currently applied data analysis techniques. His latest research has looked at predicting changes in protein activity by identifying the underlying 'Biophysical Conditional Random Field'.

FUNDING

National Science Foundation (NSF)

BIO

William Ray received his BS in Mathematics from The Ohio State University before studying an MS in Computer and Information Science and PhD in Biophysics. He currently works as the Director of the Computational Biology and Bioinformatics Division there too, as well as Associate Professor of Pediatrics at The Ohio State University College of Medicine.

CONTACT

William C Ray, PhD
Associate Professor
The Ohio State University
700 Childrens Drive
Columbus, OH 43205
USA

E: ray.29@osu.edu
T: +1 614 355 5645
W: http://www.stickwrld.org/