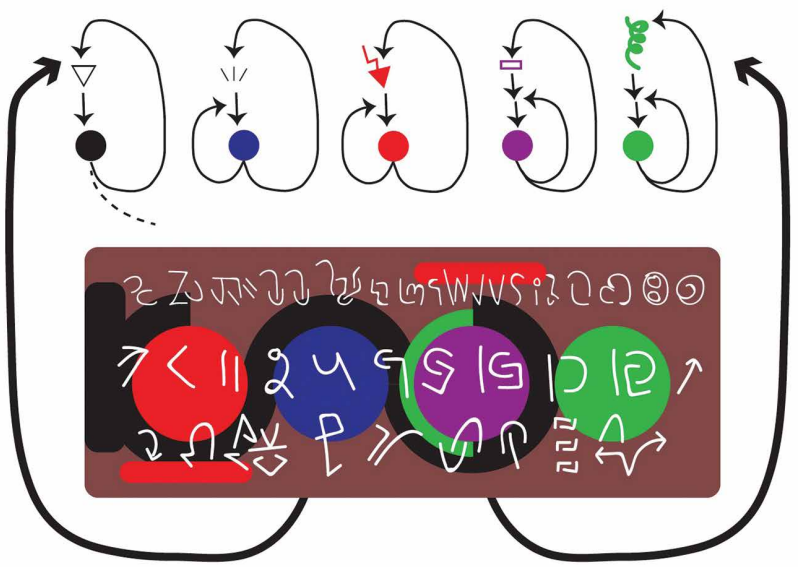
Putting Life in Context

Professor Julio Collado-Vides

Scientia

PUTTING LIFE IN CONTEXT

How do you determine important scientific links when you are flooded by new publications each day? **Professor Julio Collado-Vides** and his team at the National University of Mexico appear to have the answer.



Whether it be letters or ideograms, standardised street signs or smiley faces in instant messages, communication is one of the most important tools which humans have. The ability to share information, perhaps better known as *language*, has taken us from our first co-ordinated hunts on the savannah all the way through to present-day gossip over the Royal family.

Language is a way of conveying information, but the simple use of words by themselves is not sufficient. Simply saying 'bird' does not tell your partner whether the bird is watching them, hiding behind the bushes, cooked and ready to eat, or indeed behind them and flying away with their sandwich. Thus, words need to be linked together to provide context, using a series of rules which together make up our grammar. Despite our suffering in school, grammar is normally so well-understood for native speakers that even errors in nonsensical sentences can be identified (the classical example from Chomsky is 'colourless green ideas sleep furiously' – a sentence that makes no sense but is recognisable as acceptable grammar).

Common rules and context play a vital part in communicating information. This is not only limited to the world of human speech, but occurs in a multitude of other processes which rely on information exchange. The most fundamental of these, underlying all life itself, is that of DNA. DNA strands encode all of the information necessary for life, and yet are themselves composed of a four-letter alphabet (C,G,T,A), which codes for roughly twenty syllables (the amino acids and STOP codons) – yet this limited selection allows combination into all of the words (proteins) required to sustain life.

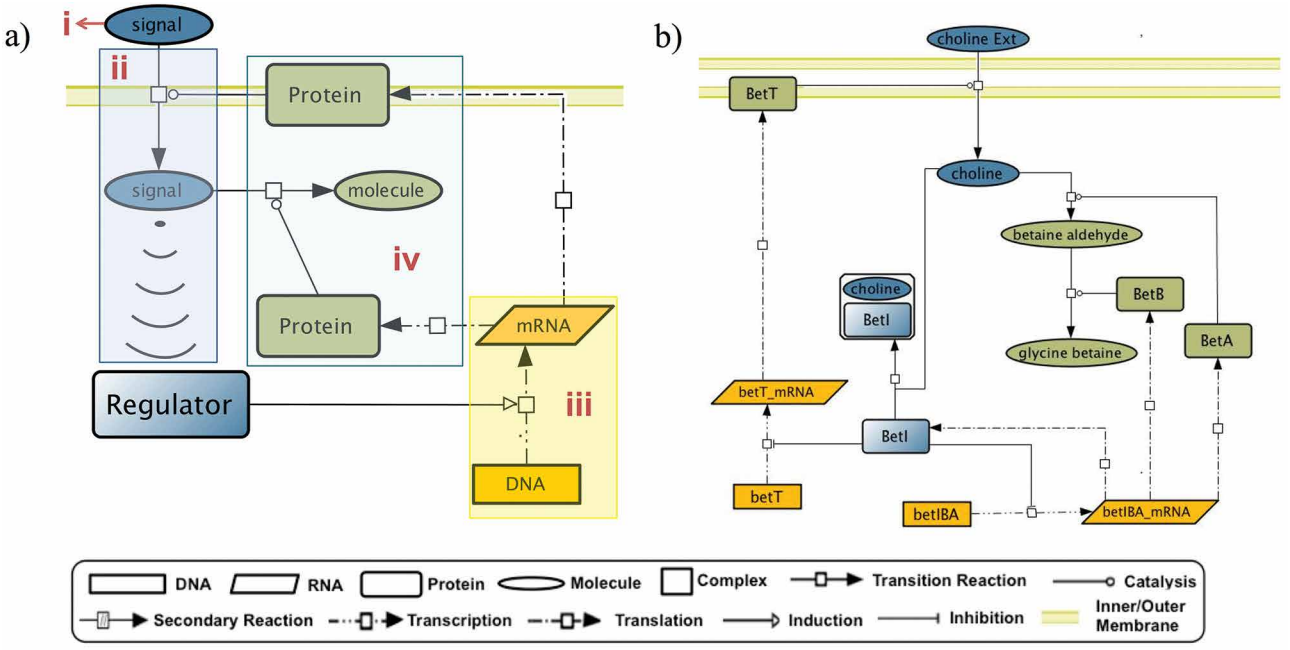
Yet, as we saw before, words are useless if you do not know the context. The scientists who finally sequenced the entire human genome were surprised to find that there were far fewer genes than they had expected. Instead the difference between, say, a skin cell and a neuron, came down to differences in context – some genes were activated in some circumstances, some in others, some were cut up and rearranged during protein production to make newly spliced variants. All of these changes were brought about as a result of genetic instructions present on the DNA strand, each of which provided part of the context required for correct understanding of the genetic 'word'. The question then is, given that DNA has words and context, can we treat it as a language?

Speaking DNA

The attempt to answer this question has long been a goal of Professor Julio Collado-Vides, of the National University of Mexico. His original studies revolved around the role of transcription factors in genetic control. Transcription factors control the first step in the production of proteins, the transcription of the DNA gene into a short-lived RNA copy. By binding to the DNA in or around the vicinity of the gene, they can control if and how well the RNA transcription machinery can interact with the gene. Although incredibly complex in practice, most transcription factors can be classified as *repressors* (which reduce the chance of a gene being transcribed) or as activators (which increase the chance).

Professor Collado-Vides' early work also involved modelling the organisation of transcription factors and gene regulation as a genetic grammar. By leveraging the significant amount of research done on language grammar, he was able to develop a model of 'transcription factor grammar' – a series of rules which allowed new transcription factor sites to be identified with far higher specificity than possible before.

‘The dream is to enable navigation of large collections of data, information and knowledge’



(a) The general GENSOR Unit concept includes 4 components: (i) signal, (ii) signal transduction shown in blue, (iii) genetic switch shown in yellow and (iv) response shown in green. (b) BetI GENSOR Unit. Example of a GENSOR Unit based on a bacterial transcription factor.

This was the starting point for over twenty years’ research in the growing field of bioinformatics – the use of computers and mathematical models to understand biological systems. This syntactic approach helped in predicting and labelling transcription factor sites in the early days of genomics – with the sequencing of the entire *E.coli* bacterial genome. Professor Collado-Vides is still highly active in the bioinformatics area, and lately he has focused on extending the initial genetic grammar to a more complex set of data – the GENSOR Units.

Mapping GENSORs

Transcription factors are a well-studied part of the genetic molecular machinery, dating back to the discovery of the bacterial LacI transcription factor in 1959. LacI acts as a repressor – binding to a strand of DNA and preventing the nearby genes from being transcribed. The repressor will detach from the DNA strand only when the correct environmental conditions are achieved – it thus acts as a link between the environment and the genetic response.

In the case of LacI, an environmental signal (lactose) is transduced into a signal that can directly affect the genetic regulator (here through the transformation of lactose into allolactose), which is then responsible for a switch in the genetic environment (the expression of lactose metabolising genes), and then leads to a response (the cell can now effectively metabolise lactose).

This combination of elements – signal, signal transduction, genetic switch and response – can be linked together by researchers, and considered as a single higher-level unit. This unit is referred to by Professor Collado-Vides and his group as a Genetic Sensory Response Unit, abbreviated to the catchier name of GENSOR Unit. Using GENSOR Units helps humans to quickly capture the incredible complexity of cell signalling, which can involve tens or hundreds of individual components, and allows comparatively simple networks of information flow to be constructed. ‘I see GENSOR Units as either a higher level of integration after concepts like operon and regulon, as a tool for understanding when analysing global genomic experiments, and as a first step to systematically describe fluxes of regulated information,’ says Professor Collado-Vides.

He and his collaborators attempted to construct a comprehensive GENSOR map of the bacteria *Escherichia coli* – known around the world to be the trusty workhorse of the microbiology lab. Their previous efforts had netted a collection of 189 transcription factors, each with differing targets, responding to different signals, and leading to different effects. The challenge, now, is to put these transcription factors together into a clear and useable network.

The group used an approach that leveraged a number of different databases, based on their experience in bioinformatics approaches. The transcription factor proteins were used as a starting point, and steadily widening automated database searches were used to find interacting proteins, known genes that were controlled by the process, and the cellular effect of those genes. The resulting pile of data was then sieved to identify the most important links. In total, 189 GENSOR elements were constructed. Given our current knowledge, only 89 are a complete genetic sensor and response systems, while for others, this approach helps in predicting some of the missing components.

The majority of these GENSOR Units were found to be controlled by some form of feedback loop – the activity of the GENSOR

is affected by its own output. These loops were predominantly simple, though some transcription factors were shown to be controlled through multi-step metabolic changes. Alongside these feedback loops were a number of different metabolic factors that either directly controlled or were directly controlled by the GENSOR Unit. In other words, a single GENSOR Unit may control the production of several different molecules, but is itself only affected by one of these.

The true advantage of GENSORs, however, lies in the ability to merge several individual units into a larger network. For example, *E. coli* have a set preference that they will follow when choosing which carbohydrate to use as a food source. Glucose is used first, then lactose, and only later come molecules such as arabinose and xylose.
A number of transcription factors, and thus GENSOR Units, are involved in carbohydrate metabolism, and the interactions of these can be used as to model just how the bacteria will behave when faced with any combination of deliciously edible sugars.

Before genomics, microbiologists were devoted to studying defined capabilities of cells, like carbon degradation, nitrogen assimilation, cell division, or responses to different environmental stresses. Transcription factors were named according to those cellular capabilities. With the advent of genomics, we now know that there is a lot of cellular integration. 'There is no elementary sensing', as Dr Collado puts it, because most molecules have multiple consequences in the cell. The GENSOR Unit is a concept adequate to describe the interconnections supporting this integrative physiology in explicit diagrams in databases.

The Artificial Librarian

What works for genetic grammar can also work for normal language, and so it is perhaps natural that Professor Collado-Vides' team would apply their knowledge to scientific communication as well. The group has long been involved in the manual collection and curation of papers covering bacterial gene regulation, and their work underlies several open databases (see for instance: http://regulondb.ccg.unam.mx/) that are regularly used by scientists to determine the current state of knowledge. However, manual curation has a number of limitations: reading papers is time consuming, the efforts of experts are required to adequately capture the findings, and the information that can be searched is limited by the database format.

Bioinformaticians, of course, are experts in drawing information out of large, complex data sources. The group asked themselves if they could use machine learning and computer-based curation in order to create a new and heavily cross-linked database of gene regulation. 'The dream,' says Professor Collado-Vides, 'is to generate methods that will impact the whole domain of people devoted to gather, organise, integrate and enable navigation of large corpora of data, information and knowledge in the biomedical sciences.'

The first step towards this dream is the integration of data mining and text analysis into several of the databases that the team is participating in. Several tools have been developed that specialise in extracting information from life science publications – they use a knowledge of how English sentences are structured to extract an overview of the results and what links have been demonstrated. These were built upon using similarity algorithms that determine 'similar' words – a search does not need to bring up an exact match to be relevant, rather a close match will also be correct. These were then integrated into an interface system that allows users to see and decide the correct piece of knowledge based on what the computer is proposing – it will highlight words and sentences it believes are involved in gene regulation and display how they link to one another.

The end result of all this work is an intelligent database that curators can use. As they read through a paper, the system automatically provides links to other publications dealing with the same subject, links determined entirely by the artificial intelligence behind the database. This means that curators no longer have to spend their valuable time hunting down correlating publications or references. Although the final process is still manual, the research team has determined that the system can increase curation speed, but most important, it will enhance the traceability of pieces of knowledge linking them to their original publication, and will enrich databases in novel ways.

Science Communication Outside the University

In yet another application of these curation ideas, Dr Collado-Vides has embarked in a non-profit adventure of what is expected to become an interactive encyclopaedia, where knowledge is organised both within an ontology (similar to the organisation of knowledge within the British Encyclopaedia), but also ordered by different levels of understanding, linking texts for laymen with more advanced texts. Conogasi (http://conogasi.org/) will initiate its activities the fall of 2017.

The Context of the Flood

In the modern genomic era, high-throughput sequencing and the expansion of research to laboratories across the world provide us with a flood of genetic knowledge as never before. However, the sheer pressure of information prevents us from understanding or even reading it all, human minds are simply not up to the task. Instead, machine learning systems allow us to simplify the flow and pull out the most useful links, be it from scientific publications or gene regulation databases. Both of these require that we understand the context of the information, the details, genetic or written, which allow us to make sense of what we see.

It is in solving this problem that the work of Professor Collado-Vides and his group truly shines. By helping to automate the recognition of context, they provide a means for us to tame and understand the flood of information. This, in turn, means that other scientists can work effectively and spend time on their true calling, the extension of human knowledge.

Meet the researcher

Professor Julio Collado-Vides
Center for Genomic Sciences
National Autonomous University of Mexico
Cuernavaca
Mexico

Professor Julio Collado-Vides received his MSc in Physical Chemistry in 1985 and went on to achieve a PhD in Biomedical Research in 1989 from the world-renowned National Autonomous University of Mexico. After his PhD, he went on to do three years of postdoctoral research at MIT. He is currently a Professor of Computational Genomics in the Center for Genomic Sciences at the National Autonomous University of Mexico. With a research career in bioinformatics and genetics spanning over two decades, he has published over 100 papers, and has been cited over 21,000 times – leading him to be recognised as one of the most highly-cited scientists in the world. He has supervised almost 20 students, sits on numerous boards, and has been awarded a number of honours – a number which can only increase. His research has been a team effort of current and past members of his laboratory (http://www.ccg.unam.mx/en/ComputationalGenomics).

CONTACT

E: collado@ccg.unam.mx
T: (+52) 777 313 9877
W: http://www.ccg.unam.mx/en/personal/julio-collado-vides

KEY COLLABORATORS

Daniela Ledezma-Tejeida (PhD student working on the GENSORs)
David Rosenblueth at IIMAS, UNAM (computer scientist who implemented the programs of the grammatical model http://turing.iimas.unam.mx/~drosenbl/)
Jacques van Helden, Université d'Aix-Marseille, France (http://jacques.van-helden.perso.luminy.univ-amu.fr/)
Fabio Rinaldi, Swiss Institute of Bioinformatics (http://www.sib.swiss/rinaldi-fabio/rinaldi-fabio-sub)

FUNDING

UNAM
CONACyT Mexico
NIH (NIGMS)

REFERENCES

D Ledezma-Tejeida, C Ishida, J Collado-Vides, Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in Escherichia coli, Frontiers in Microbiology, 8, 1466. DOI: 10.3389/fmicb.2017.01466

S Gama-Castro, H Salgado, A Santos-Zavaleta, D Ledezma-Tejeida, L Muñiz-Rascado, JS García-Sotelo, K Alquicira-Hernández, I Martínez-Flores, L Pannier, JA Castro-Mondragón, A Medina-Rivera, H Solano-Lira, C Bonavides-Martínez, E Pérez-Rueda, S Alquicira-Hernández, L Porrón-Sotelo, A López-Fuentes, A Hernández-Koutoucheva, VD Moral-Chávez, F Rinaldi, J Collado-Vides, RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, Nucleic Acids Research, 2016, 44, D133–43.

S Gama-Castro, F Rinaldi, A López-Fuentes, YI Balderas-Martínez, S Clematide, TR Ellendorff, A Santos-Zavaleta, H Marques-Madeira, J Collado-Vides, Assisted curation of regulatory interactions and growth conditions of OxyR in *E. coli* K-12, Database (Oxford), 2014, pii: bau049.

J Collado-Vides, Grammatical model of the regulation of gene expression, Proceedings of the National Academy of Sciences of the USA, 1992, 89, 9405–9409.

UNAM