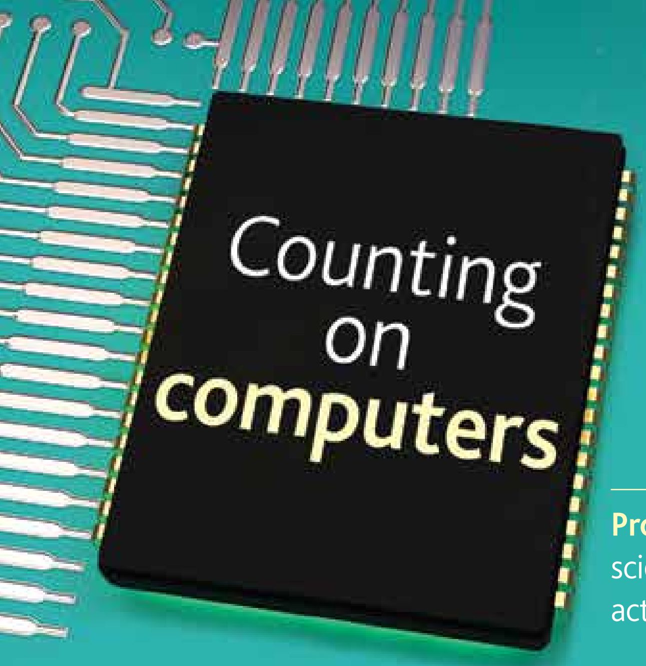


Counting on computers

Professor Stefano Toppo reveals how the evolution of computer sciences has enabled his group to contribute to a wide variety of activities that may usher in a new phase for protein function prediction

Can you describe your professional background? What inspired you to create a system that could predict protein function on the genomic scale?

I've worked in molecular biology and whole genome shotgun projects since 1994 in which I was involved in DNA library construction and sequencing. During this period, I shifted my interests from laboratory life to software development, applying computer science to the data analysis of '-omics' projects. In 2002, I got into biochemistry with a research assistant position through which I became acquainted with the protein world, their structure and function, both from an experimental and computing perspective. The knowledge I gained in these various research areas led me to develop an interest in predicting protein function at a genomic scale, taking into account the expertise acquired both in genome and proteome contexts.

Could you define 'bioinformatics'?

Bioinformatics is a broad discipline that gathers many different fields of research. A neologism coined in 1970 by theoretical biologists Paulien Hogeweg and Ben Hasper, bioinformatics embraces any aspect of data analysis of biological experiments involving DNA and protein sequences. It is placed in the middle between computer science and biology.

What are the main topics of research being carried out in the University of Padua's Department of Molecular Medicine?

Our department covers many aspects of biomedicine from microbiology, virology, biochemistry, genomics, proteomics, advanced molecular and cell biology in cancer research up to translational medicine and clinical research. It's a complete and complex department, and as group leader I'm in charge of supporting bioinformatics analyses for all of the different lab groups. We have been involved in many research projects in virology, microbiology and biochemistry through which I've had the chance to hone my expertise by studying the details of protein families at the molecular and sub-molecular levels using computational techniques.

You have worked on a computer program called 'Argot'. Can you explain what this tool does, how it operates and how it is improving?

Argot stands for 'Annotation retrieval of gene ontology terms'. The working principle appears simple but, in reality, its complex algorithm exploits various strategies to investigate public protein databanks. It's a sort of data miner which aims to discover useful knowledge in public databases that can be further processed to attribute a function to the proteins of interest. Argot2 is mainly a refinement of our original idea but this has allowed us to rank second overall in a recent international challenge. Even still, Argot2 has a lot of room for improvement; for example, we plan to extend the concept of function using some transitive properties that can link distantly related proteins in a manner that, as far as we know, has never been attempted before. Our approach could help understand still unknown molecular mechanisms at the basis, poorly studied organisms and human cell functions.

How important is collaboration to your endeavours?

I must absolutely mention the most important of my colleagues, Dr Paolo Fontana, who works as Group Leader at the computational centre of the Edmund Mach Foundation near Trento, Italy. Without his contribution Argot would never have seen the light. I would also like to mention the components of my small group: computer scientist Dr Marco Falda, and bioinformatics biotechnologist Dr Enrico Lavezzo.

What has been the biggest challenge and how did you overcome it? Has this opened up new avenues for investigation?

The biggest challenges are hidden in seemingly simple questions that come to me from the people who work in my department. My job is to transform requests into something doable from a computational point of view. This is the task of bioinformatics business: trying to find solutions that straddle disciplines, using different languages and having different needs, such as computer science and biology. Recently, for example, I've been asked to create a set of short DNA fragments, called primers, in order to barcode exogenous DNA fragments that integrate randomly in the human genome as viral vectors and retrieve their exact position in the genome at a later stage. To do that, one needs to create small tags of DNA (a string of 20 nucleotides) that are absent in the genome; easy to say but almost impossible to achieve. We have solved the problem using computer graphics cards and an efficient algorithm that has never been attempted before. Patenting these tags has potentially opened up new and interesting scenarios that I think can help dramatically with recent techniques in genetic engineering.

Protein prediction

Despite major computational advances in whole genome sequencing, existing tools are sagging under the volumes of data being produced. Pioneering strategies at the **University of Padua** might finally be able to handle the load

WHOLE GENOME SEQUENCING (WGS), which determines the full DNA sequence of an organism's genome, is a rapidly expanding field that is increasingly becoming a standard procedure in laboratories across the world. With manual annotation now entirely impractical for cutting through such vast swaths of data, bioinformatics is evolving fast alongside WGS as more efficient *in silico* methods of protein function prediction become a necessity. In the midst of this boom in the availability of genomic information, the means for understanding the fundamentals of how organisms develop and function have never been so widespread. Yet, despite such advances, even for the most studied organisms, current knowledge regarding the function of a large number of gene products is woefully limited.

Simply knowing a protein function does not count for much on its own. To understand what it does within the cell and achieve a more profound understanding of life at the molecular level, a full annotation must take into account where, when and how a protein performs its role. Existing computational methods of protein function prediction might be reaping the benefits of recent advances in processing capacities, but there is an urgent need for greater accuracy and reliability across the board. Within this restless field, scientists are developing a pioneering new generation of tools that could take genome sequencing into its next phase and hugely impact the biomedical and pharmaceutical worlds.

THE ONTOLOGY OF GENES

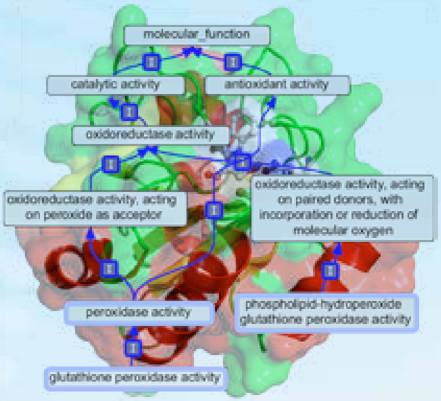
As Professor of Biochemistry at the University of Padua's Department of Molecular Medicine, Dr Stefano Toppo is currently engaged as Group Leader in a broad range of exciting projects with his colleagues Drs Marco Falda and Enrico Lavezzo. Among the group's rich contributions toward DNA and protein studies, their input into the field of computational protein function prediction has received extensive praise within the community and placed them at the forefront of current efforts to improve the speed and accuracy of gene annotation.

Facilitating such widespread activity in developing the next generation of tools for function prediction is the Gene Ontology (GO) initiative, a large-scale project dedicated to standardising the language of the related disciplines. With terminology freely adopted for one species and not another, a lexicon specific to a single research area (and even on occasion a single research group) embracing each other's findings with anything approaching fluency is often a huge challenge for scientific communities. A shared set of well-defined terms for gene products will not only enable faster evolution of annotation tools but should also make function discovery for newly found sequences a far easier task.

NOVEL ANNOTATION

Adopting the terminology prescribed by the GO project, Toppo's first foray into the design of protein annotation software culminated in a program called Argot: Annotation retrieval of gene ontology terms. It is a powerful tool capable of processing thousands of sequences and making speedy inferences about protein function with highly promising levels of specificity and sensitivity. Developed with Dr Paolo Fontana, the success of Argot's performance can be explained by a novel approach that assembles GO terms according to their calculated weights and semantic similarity. Even though it can carry out both large and small operations equally well (with a large-scale genome project being just a few hours' work on a standard desktop computer), Argot is just the first deliverable in an ongoing programme of development.

Still using the principal ideas at the project's core, Argot2 runs on a completely rewritten engine that has resulted in marked improvements upon the efficiency of its forebear. Toppo states: "We have boosted computation without affecting the precision of our tool in prediction accuracy". With its entirely revamped process of annotation, Argot2's initial performance trials on grapevine and apple genomes put it in good stead before an evaluation of the community's efforts in protein function prediction at the Critical Assessment of Function Annotation (CAFA) experiment. With 30 teams from 23 research groups, CAFA's review took in 54 methods of annotation over a 15-month period that tested the tools on a target set of proteins from 11 organisms. The overhaul given to Argot outpaced the vast majority of methods, ensuring the newer model's high ranking in second place. However, CAFA's review has revealed an unfortunate trend amongst even the best annotation tools: while all 54 efforts were improvements upon the first generation, with particularly accurate results in determining molecular function, all were sorely lacking in the 'biological process' category. The gulf between even the most capable tools' abilities and an acceptable level of annotation is still wide.


molecular_function
catalytic activity
antioxidant activity
oxidoreductase activity
oxidoreductase activity, acting on peroxide as acceptor
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
peroxidase activity
phospholipid-hydroperoxide glutathione peroxidase activity
glutathione peroxidase activity

A gene ontology molecular function graph of glutathione peroxidase.

The Intelligent Systems for Molecular Biology (ISMB) conference 2012 in Los Angeles. From left: Drs Enrico Lavezzo, Marco Falda, Stefano Toppo and Paolo Fontana.

CLOSING THE GAP

Attempting to traverse this gap, Toppo is looking at the next generation of the Argot tool from a new perspective. Protein function is inferred most commonly by the similarity between sequences. However, with sequence similarity alone, large stores of information can remain hidden in databanks. Toppo is looking to exploit this weakness by making selections based on the functional similarity between proteins, improving Argot so it can grasp the where, when and how of protein behaviour. With a functional space divided into three hierarchical layers, the next Argot tool will hold data on phylogenetically related proteins at the bottom, functionally similar proteins in the middle and models for revealing the complex metabolic and signalling pathways that current methods are not capable of at the top. The successful development of such a tool could have far-reaching implications, as Toppo explains: "Our approach could help elucidate still unknown molecular mechanisms at the basis not only of poorly studied organisms but also of human cell functions". The potential impact of the third Argot on human health may be strongly felt indeed.

EXTRACURRICULAR ENDEAVOURS

Although the Argot tool has been evolving steadily since its creation in 2008, Toppo's contributions to the wider worlds of DNA and protein research have not abated. In the aftermath of a meningitis epidemic in the Veneto region of northern Italy, the *Neisseria meningitidis* bacterium responsible for the outbreak provided an opportunity to gain interesting insights through *in silico* genome assembly. Using shotgun and paired-end sequencing strategies, Toppo discovered the isolates of *N. meningitidis* in his laboratory contained coding sequences from other serogroups, other Neisseria species and even from other bacterial species altogether. Although horizontal gene transfer (HGT) in *N. meningitidis* is already known, these unexpected features in its genome highlight a greater capacity for HGT than was previously known.

Also, the number of known isoforms of glutathione peroxidases enzymes has recently gained an extra member as a result of the research conducted in Toppo's lab, with the putatively named glutathione peroxidase eight (GPX8) bringing the total up to eight types. Of special interest, however, is the isoform GPX4. The discovery of a dormant form has enabled Toppo to shed light on processes such as neurodegeneration, tumour proliferation and metastasis through its capacity to affect the balance between cell survival and cell proliferation. Conducting consistently groundbreaking research, important discoveries like these will only be made more often as the computational disciplines and tools supporting such research continue to evolve, a progression Toppo is making strident efforts to facilitate.

There is an urgent need for greater accuracy and reliability in protein function prediction

INTELLIGENCE

MOLECULAR MEDICINE

OBJECTIVES

- To clarify how protein function is conserved in distantly related proteins and organisms, and help researchers to formulate hypothesis-driven experimental designs
- To develop an integrated framework for the annotation of poorly studied organisms, including their metabolic and signalling pathway reconstruction
- To extract knowledge from unlinked and sparse data available in public databanks and investigate unexpected correlations that can be exploited for the annotation of protein function

KEY COLLABORATORS

Dr Enrico Lavezzo; **Dr Marco Falda**, Department of Molecular Medicine, University of Padua, Italy

PARTNERS

Dr Paolo Fontana, Edmund Mach Foundation, Italy

FUNDING

Local grants provided by the University of Padua

CONTACT

Professor Stefano Toppo

Department of Molecular Medicine
University of Padua
via U Bassi 58/b
35131 Padua
Italy

T +39 049827 6958
E stefano.toppo@unipd.it

www.medcomp.medicina.unipd.it/stefano

STEFANO TOPPO is Professor of Biochemistry at the University of Padua, Italy. He began his studies in molecular biology and bioinformatics and obtained a BSc in Biology and PhD in Biotechnology in 1994 and 2001, respectively. As Group Leader he investigates genome-scale, high throughput sequencing projects, genome assembly and protein function prediction. He is also expert in the study of specific protein families, such as glutathione peroxidases.

ARGOT

DMM
Dipartimento di Medicina Molecolare
Università di Padova