It may not be entirely an alarmist view that we are living somewhat dangerously these days. The possibility of catastrophic disease outbreaks remains a clear and present danger for each and every one of us. While we may have managed to avert disastrous outcomes with the recent close shaves with SARS and avian flu, it is important to remember that these diseases have not been eradicated and could unexpectedly re-emerge as the viruses mutate and return with a vengeance. In addition to having to deal with the increasing number of new emergent diseases, we are also facing the returning threats from diseases previously regarded as conquered. For example, common bacteria such as Staphylococcus aureus, Serratia marcescens, and Enterococcus sp. have developed resistance to the strongest available antibiotics, and are now colonizing and causing disease in the general population. Cases of tuberculosis resistant to all traditionally effective treatments have also recently emerged to the great concern of health professionals.

It is obvious that we will be fighting a losing battle if we continue with the traditional ways of vaccine and drug development, which take decades of effort and require billions of dollars as well as probably a good dose of serendipity. To stand a fighting chance against the new health threats, it is crucial to drastically reduce the cost (in terms of time and dollars) of developing vaccines and drugs. This may be possible if we arm ourselves with a wealth of data and the power to turn the data into useful knowledge.

The rapid acquisition of large-scale biological data has become a reality with recent advances in biotechnology that make it possible to read out the blueprints (i.e. DNA) of various organisms (including human and the various disease-causing viruses and bacteria; see Box 1), as well as other biological data such as the quantity of various biomolecules in different cells (see Box 2 for the various significant biotech innovations generating biological data). The consequence of this recent proliferation of high-throughput methods in molecular biology is the spawning of an unprecedented data explosion that cannot be handled in the traditional fashion. Data mining — the ability to find unexpected patterns in large sets of accumulated data — has become instrumental in the discovery of useful biological knowledge in the postgenome era, enlisting a new breed of computer scientists (or rather, bioinformaticians) who can equally understand the technicality of computational algorithms and the complexity of biological data.

Located at the corner of the National University of Singapore (NUS) campus, the Institute for Infocomm Research (I2R) — a member of the Agency for Science, Technology and Research (A*STAR) — has been steadily
building up its expertise in informatics and harnessing its power to tackle biological and medical problems. The 20-plus–member team at the Knowledge Discovery Department there has been working tirelessly to devise novel data mining algorithms in order to help biologists harvest the benefits of the current deluge of biomedical data, joining the biologists to combat hand in hand the many health threats that we face today.

**Speeding Up Vaccine Design Computationally**

As mentioned earlier, the importance of pandemic preparedness can no longer be overlooked given the recent emergence of various life-threatening viruses such as SARS and the avian flu. Infectious diseases, such as tuberculosis, malaria, and HIV/AIDS, cause around 25% of global deaths. Moreover, the levels of resistance to antibiotics by many bacteria keep increasing.

Vaccines help prevent morbidity and mortality due to infectious diseases, and are among the greatest public health successes. Vaccine research has grown to be an important research field with global annual R&D funding of around US$5 billion. Finding the right protein (i.e. “target”) to make a vaccine with wide population coverage for a disease is a difficult task. As finding vaccine targets among a host of proteins is akin to searching for a needle in a haystack, a computational target prediction system that is accurate and broadly applicable would be extremely helpful for vaccine designers.

Due to the high polymorphism of human immune cells, the response to foreign substances (e.g. viruses, bacteria, cancerous cells, etc.) varies from person to person. The selection of suitable protein peptides for designing vaccines with broad population coverage is a challenging process, as only 1%–5% of the peptides are suitable candidates.

To overcome this major hurdle, the I2R team has developed state-of-the-art computer-aided vaccine design (CAVD) technologies that can intelligently predict promiscuous peptides for broad coverage. In addition, the prediction algorithms must also overcome inherent data challenges such as noisy data and the lack of experimental data. Various classification methods, such as quantitative matrices, artificial neural networks, hidden Markov models, support vector machines, and genetic algorithms, have to be optimized before being employed as predictive engines in the system.

By working closely with bench biologists, the I2R team has recently applied its CAVD technologies on various viruses such as the dengue, West Nile, yellow fever, and human influenza A viruses. The prediction system was found to achieve up to 89% accuracy in the computational identification of immunological hotspots (the equivalent of vaccine targets) in the dengue and West Nile viruses: nine hotspots were predicted for these two viruses by the I2R team using purely computational algorithms, and eight were subsequently
experimentally validated by scientists in the US. This translates into a significant reduction of 70%–90% in experimental screening, and is thus instrumental for accelerating the development of effective vaccines as required by the current scenarios. By being able to take advantage of the recent deluge of new information from the genome projects and turning them into useful knowledge, our CAVD technologies (Fig. 1) can therefore become a formidable tool for vaccine researchers and pharmaceutical companies to accelerate their search for vaccine targets in their combat against the various viruses and other infectious diseases.

**Fig. 1.** Computer-aided vaccine design (CAVD) at I²R. We combine the computational methods in the dry lab with the knowledge from our bench biologists in the wet lab to develop various web-based prediction tools for scientists worldwide in predicting vaccine targets (see Box 3 for a detailed listing of the web tools).

---

Aiding the Design of Drugs Computationally

Vaccines serve as protection or prevention against contracting diseases, while drugs are medication for treating diseases (either to cure a disease or ameliorate the harm from the disease). Unlike a vaccine, a drug is typically a biological compound that can be synthetic or nonsynthetic, and it is synthesized externally of an organism to be effected internally of the organism.

However, the discovery and development of new drugs for the treatment of diseases is also a costly and time-consuming process. It is estimated that only 1 in 5000 compounds or fewer is approved by the US Food and Drug Administration (FDA). The total duration from initial discovery to commercialization of a new drug averages around 15 years with estimated costs of up to US$750 million. The team at I²R is seeking to develop effective computational models to help reduce the number of experiments, so that researchers can concentrate on biological assays in the laboratory on the most promising components.
The main challenge of rational drug design is the identification of compounds that can bind to a chosen target for a particular disease. This is a multifaceted problem. First, it requires knowledge of the active site in the drug target — no information exists for more than 90% of protein sequences, and sequence and structure information are growing at an exponential rate. Although the Protein Data Bank (PDB) doubles every three years and the sequence data banks double every 17 months, the majority of these sequences represent novel protein families with unknown function, and limited success has been achieved for the computational prediction of receptor active sites. Secondly, it has to deal with protein flexibility that often requires a large number (in the hundreds or thousands) of degrees of freedom: each structure undergoes numerous essential minor and major reorganizations upon binding to other proteins or chemical substrates. Thirdly, it requires the screening of massive libraries of compounds for their ability to inhibit or stimulate the target — a nontrivial computational problem that may take several man-months to man-years on a single computer. A new generation of enabling technologies to address the above is needed for the drug discovery market.

The *in silico* drug discovery team at FR has embarked on devising efficient solutions to these challenges in developing a computational system for identifying possible drug targets by providing solutions for rapid high-throughput and cross-screening of massive libraries of compounds (Fig. 2). The methods used in the system have since been filed for patents, and the team is currently collaborating with several partners to further develop these methods.

---

**Fig. 2.** Summary of virtual screening protocol for *in silico* drug discovery as implemented by the FR team.
Bioinformatics in Asia Pacific

Computational Challenges

For centuries, biology has been an empirical field featuring mostly specimens and petri dishes. With the advent of high-throughput biotechnologies, however, modern molecular biology and medical research now involves enormous amounts of data of increasing variety and complexity. The use of informatics to organize, manage, and analyze these data has clearly become an important element of biology and medical research. The bioinformatics team at I2R has devised novel data mining algorithms, and has successfully applied them on various types of biomedical data to discover new and useful knowledge, such as better treatment plans from cancer microarray data and new protein complexes from whole-genome protein interaction networks.

As described above, the team will continue to employ data mining in joining the biologists in the urgent combat against the many health threats that we face today by enabling the necessary speed-up for the design of vaccines and drugs through intelligent computational approaches.

The Long Road to Reading the Blueprints of Life

1865 — Gregor Mendel discovers the laws of genetics.
1953 — James Watson and Francis Crick describe the double-helical structure of DNA.
1977 — Frederik Sanger, Allan Maxam, and Walter Gilbert pioneer DNA sequencing.
1982 — The US National Institutes of Health establishes GenBank, an international clearinghouse for all publicly available genetic sequence data.
1985 — Kary Mullis invents polymerase chain reaction (PCR) for DNA amplification.
1985 — Leroy Hood develops the first automatic DNA sequencing machine.
1990 — The Human Genome Project begins, with the goal of sequencing human and model organism genomes.
1999 — The first human chromosome sequence is published.
2001 — A draft version of the human genome sequence is published.
2002 — A physical map of the mouse genome is published in Nature.
2003 — The Human Genome Project ends with the completed version of the human genome sequence.
2003 — Canadian scientists finish mapping the genetic sequence of a coronavirus.
2004 — The first draft of the chicken genome assembly is published. Analysis of the chicken sequence involves an international group of scientists from the US, UK, Europe, and China.
2006 — 451 complete genomes have been published, and the number is growing (see www.genomesonline.org).