Proteomic Progress: Bangalore Institute Becomes a Major Center of Biotech Research

Chandra Shekhar
DOI 10.1016/j.chembiol.2010.03.003

In the U.S. or Europe, India is often associated with information technology or service outsourcing, not with scientific research. Nonetheless, largely unnoticed by western media, India has been making strides toward becoming a research powerhouse in the life sciences. Enjoying unique manpower and cost advantages, this country now has several major universities and scientific institutions with active programs in this area. Spurred by supportive government agencies, these institutions are producing an increasing number of publications, patents, and industry spin-offs. Consider Bangalore-based Institute of Bioinformatics (IOB), a relatively small 40 member outfit that has nonetheless become a major global player in proteomics research. In less than a decade of existence, the institute has produced what is arguably the world's best curated protein database and is well on its way to replicating this success in cancer biomarkers, proteogenomics, and signaling pathways. "We undertake projects that most other labs would find extremely hard," says IOB's founder and director Akhilesh Pandey, M.D., Ph.D., who is also a researcher at Johns Hopkins University School of Medicine. "And we accomplish them in a time frame that would be virtually unthinkable elsewhere."

Proteogenomics, the use of proteomics to annotate genes, is an area where IOB has carved out a niche for itself.

This is not a vain boast. Take, for instance, the Institute's Human Protein Reference Database, an unprecedented online compendium of curated protein information. Containing information about more than 27,000 proteins and 39,000 protein-protein interactions, the database is the fruit of several years of effort from curators who sifted through more than 2,000,000 research papers. Unlike most other protein databases, IOB's creation encapsulates virtually every relevant feature of each protein—function, sequence, domains, motifs, interactions, expression, localization, modifications, disease associations—and includes results obtained with almost any experimental platform. Initiated in 2003, this resource is now in wide use: among its many clients are the biological network visualization tool Cerebral, the sequence analysis tool CompariMotif, and the type 1 diabetes research database T1Dbase. The website that hosts this resource now gets millions of hits each month. The database is also the scaffold for the institute's Proteinpedia, a unique repository of contributed proteomic datasets that includes nearly two million peptides from about 2,700 experiments from 75 laboratories worldwide. Proteinpedia too has become a valuable resource for proteomics. "Researchers across the world recognize and appreciate us as human protein database people," says Keshava Prasad, Ph.D., an IOB faculty scientist who coordinates this database today.

Pandey conceived the idea of a Bangalore-based bioinformatics institute nearly a decade ago. A proteomics researcher, Pandey was an early adopter of high-throughput techniques such as mass spectrometry. In fact, as a visiting scientist at Matthias Mann's laboratory, then at the University of Southern Denmark, he developed SILAC (Stable Isotope Labeling with Amino acids in Cell culture), a revolutionary technique to observe a cell's changing proteome. The problem with such techniques, Pandey observed, was that interpreting the vast amount of data they generate needs a large team of biologists skilled in computational methods. In the U.S., finding such a team can be very hard and funding them, even harder; "That's when I thought of Bangalore," says Pandey. An institute based there could recruit and train biologists to excel in the painstaking, intensive, and often laborious tasks that high-throughput proteomics requires.

Most of Pandey's friends and colleagues thought the idea was crazy. Even those who saw merit in it, such as Pandey's erstwhile mentor Mann, wondered, "But where is the money?" But such skepticism only served to make Pandey even more determined. By spending all his savings, then borrowing to the limit of his credits cards and finally borrowing from his brother, Pandey managed to get IOB up and running in December 2002. "When everyone opposes you, the chances are you are doing something right," Pandey jokes. It wasn't until much later that grants from various U.S. and Indian sources helped ease the funding pressure.

That was only one part of the institute's struggle. "Now we are seeing the good times, but we've had to come a long way," recalls Harsha Gowda, Ph.D., who joined the institute soon after its inception in 2002. At that time, apart from dealing with equipment, infrastructure, and staffing issues, IOB had to struggle to establish its reputation. As a research start-up, it had no track record yet; its unconventional structure as a privately funded research institute made it a hard sell in the eyes of funding agencies. The turning point came when some of its initial research efforts culminated in prestigious publications. One of these was the annotation of the human X chromosome. At that time, the lab that sequenced the chromosome was engaged in a similar effort. "As a small, non-sequencing center, we were fighting against the odds," says Gowda. Nonetheless, both efforts were rewarded with papers in the April 2005 issue of *Nature Genetics*, accompanied by an

editorial that praised IOB's achievement as "a feat worth replicating." As a research institute, IOB had "arrived." A string of other publications on protein databases, cancer biomarkers, and other topics consolidated its reputation. Grant money flowed in more readily. "In fact, we have had 100% success in the past few grants we've applied for," says Gowda. "Now IOB is accepted as one of the premier research institutes in India."

Proteogenomics, the use of proteomics to annotate genes, is an area in which IOB has carved out a niche for itself. The motivation is simple: genome sequencing has become increasingly faster and cheaper, but genome annotation—identifying and characterizing genes—has lagged behind. Given a nucleotide sequence, researchers typically rely on algorithms that predict genes either directly or by homology with known genes from similar organisms. Neither method is foolproof. Most existing genome annotations are riddled with errors such as missing, split, or truncated genes; wrong start codons or wrong N termini; and short, noncoding regions mistaken for true genes—so much so that almost half the genes in a typical genome annotation are labeled as "hypothetical" unless confirmed by other methods. "To annotate genes reliably, you must understand the three 'omes': genome, transcriptome, and proteome, and you should have the instrumentation, the bioinformatics capability, and the inclination for it," says Pandey. "At this point in time, there's only IOB which can pull it off."

IOB's work on the malaria-carrying *Anopheles gambiae* mosquito demonstrates the utility of its approach. The existing genome annotation listed about 13,000 genes. By carrying out a proteomic analysis of certain organs of the insect, IOB researchers identified 8,675 unique peptides. Of these, 94 lay inside introns, 5 lay in UTRs, 12 overlapped intron-exon junction, 42 lay near mapped genes, and so on: in total, such anomalies helped correct nearly 200 gene annotations. "The most thrilling part was that we identified and validated around 35 novel genes," says Sutopa Dwivedi, a doctoral student who plans to annotate the related *A. stephensi* mosquito next.

Other IOB candidates in various stages of proteogenomic annotation include the leishmania parasite; the tuberculosis bacterium; *E. coli*; the tomato, mango, silkworm, basil, and neem plants; and several species of yeast. Pandey hopes to expand these efforts and establish a center of proteogenomics that will be a first of its kind in the world. "We believe that in the future, when scientists sequence a new genome, they will also sequence its proteome and then put out the data in the community," says Pandey. "This will become the only acceptable way to do things."

Another ambitious ongoing project at IOB is NetPath, a curated database of human signaling and metabolic pathway information. A vast amount of information about cellular signaling events is available in the literature. A database that collects and organizes this widely scattered information in the form of pathways would be of immense value to systems biology studies. This task, however, is complex, intensive, laborious, and can't be automated; as a result, the pathway resources that exist have many limitations and none are fully curated. As with the human protein reference database and other large-scale curated resources, IOB has again stepped forward to build a "one-stop shop" for human pathways. Several university students are being trained to work alongside IOB researchers to complete this mammoth task, says Kumaran Kandasamy, who leads the curation effort. A preliminary version is already up and running, with 10 immune signaling pathways; IOB hopes to expand it to about 500 signaling and 500 metabolic pathways. "We have a clear goal in the coming three years: to become the number one source of all pathway information in humans," declares Pandey.

While pursuing this and other ambitious goals, IOB has to deal with many challenges. Pandey may no longer need to max out his credit card to pay the institute's bills, but funding remains a concern. The institute's annual running cost of about $350,000, though small by U.S. standards, is still a considerable amount. Generous grants from various Indian funding agencies notwithstanding, Pandey and colleagues can't afford to spend too lavishly. For instance, "although we have state-of-the art equipment, we can't yet afford a cold room," says Pandey, pointing to the refrigerators that now serve the purpose. Many conveniences taken for granted in the West are harder to obtain in a third-world country. Key reagents may take several weeks to arrive after ordering; this precludes a trial-and-error approach to find the best pipelines for various types of experiments. "We can't afford to play around too much here," says Pandey. "That's why we prefer to work with well-established pipelines." Although infrastructure issues have been minimized, some remain. Equipment failure is a potential concern, as replacement parts may be hard to find. Internet speeds in India remain modest compared to the West; IOB's protein databases are hosted on U.S. servers to ensure a fast response.

Despite these challenges, IOB's rise as a proteomics research center has been nothing short of meteoric. Armed with the equipment, manpower, skills, and willingness to take on large-scale proteomics projects, this tiny Bangalore institute is now poised to "play with the big boys," says Pandey. Indeed, so pleased is he with its success in proteogenomics that he is considering setting up a similar center in the U.S.. "An Indian institute becoming a model for the West, that would have been unthinkable before," he says. "But we are beginning to permanently change the way things are done."

Chandra Shekhar (chandra@nasw.org) is a science writer based in Princeton, NJ.

