Profile of Janet M. Thornton

Although many scientific disciplines have experienced vast growth over the past three decades, few may have come as far as bioinformatics. When Janet Thornton, a structural and computational biologist, first began analyzing protein structures in 1973, the structures for only a handful of proteins had been solved, and the genome for even the smallest virus was unknown. Today, the number of solved protein structures has grown to over 30,000, and complete sequences for dozens of organisms have been decoded. "We went from having virtually nothing to having almost the whole biological spectrum of protein types," says Thornton.

Director of the European Bioinformatics Institute (EBI) in Cambridge, England, Thornton has led the charge throughout the bioinformatics boom, especially as it relates to protein structure. In addition to her position at EBI, she is also coordinator of the BioSapiens project, which aims to address the current fragmentation of European bioinformatics by creating a virtual research institute.

Thornton's main research goal has been to tie together the relationships between protein sequence and structure and, more recently, between structure and function. "We're now at the stage of trying to understand how protein function evolves, either in different organisms or cell types, to ultimately create life," she says. Along the way, Thornton has helped design myriad computer programs and databases to analyze protein sequence and structure data. In her Inaugural Article in this issue of PNAS (1), she presents a computer application that can predict functional similarities among proteins in a given family by comparing key functional residues.

Thornton has received numerous awards and honors, including being named as a Commander of the British Empire, Fellow of the Royal Society, and Honorary Professor at Cambridge University (Cambridge, England). In 2003, she was elected to the National Academy of Sciences as a Foreign Associate.

Sibling Rivalry

Thornton first developed a curiosity about nature and the world around her early on in life. Growing up, she enjoyed taking nature walks along the English coast and became interested in fields such as geology and astronomy. "I really enjoyed biology as well, but at that stage biology was more of a descriptive science and not an understanding science," she explains. "I think I was drawn to physics because it was all about how you could describe things and make equations to understand why things worked the way they did."

Janet M. Thornton

A wish to avoid comparison and competition with her older sister Margaret may have helped as well. "She was always a bit cleverer than me," Thornton jokes. So, after her sister went on to pursue a classics degree, Thornton decided the best way to avoid further sibling rivalry would be to get a degree in a field as distant from classics as possible, and, for her, physics fit the bill.

Thornton enrolled at the University of Nottingham (Nottingham, England) in 1967, but even as she was completing her degree in physics, she knew she wanted to apply her knowledge outside of pure physics. "I wasn't interested in machines, I was more interested in the natural world," she says. "So when the option came to work at the National Institute for Medical Research (NIMR) at Mill Hill, that gave me a chance to merge my physics background into a more biological research career."

At NIMR, Thornton studied the conformations of dinucleotides by using spectroscopic and computational methods (2). "I approached my project from a physics end, exploring various combinations to try and find the lowest-energy conformations," she says. "I wasn't worrying too much about what dinucleotides did functionally. I was much more focused on shape and structure." While working on her Ph.D. project, Thornton also pursued a Master's degree in biophysics at King's College in London, to help direct her research toward biology.

Thornton particularly enjoyed the computing aspects of her dinucleotide project and wrote several software programs to visualize conformations and calculate free energies (3). "But regrettably I never used them again after I left," she says, "and neither did anyone else." Still, her computing skills came in handy when she began looking for postgraduate work in 1973. "I wanted to deal with real, hard experimental data," she says, "but no jobs were advertised for people like me." Thornton eventually landed a position as both a research fellow and systems administrator in David Phillips's crystallography laboratory at Oxford University (Oxford).

One of the founding leaders of structural biology, Phillips was a generous and supportive advisor to Thornton. Her move to his group proved to be a refreshing return to a familiar world. "I found my graduate work difficult at times," says Thornton. "Mill Hill was heavy on cellular biology, and coming in as a physicist, I didn't know any of the language," she says. "And when I got to Oxford, it was wonderful because everybody was molecular again."

Physics vs. Proteins, Order vs. Chaos

At Oxford, Thornton began working in the then-emerging field of bioinformatics, so nascent that it had yet to be named. "I started out trying to understand how the protein sequence determines the three-dimensional structure," she says. "My goal was to see if one really could predict a structure based on the sequence." The process was complicated by a lack of structural data to use as a resource. "When I began, I think that only about 10 protein structures were available in the data bank."

Further complicating matters for someone accustomed to the ordered world of physics were the chaotic tendencies that proteins sometimes exhibited. Thornton noticed this messy behavior when she began studying the sequences of numerous proteins and comparing proteins from different families. "In physics, you have a set of laws, and nature obeys those laws," she says, but "simple rules describing [proteins] are elusive because of their complexity and are therefore often best captured as propensities rather than rules."

Thornton began making sense out of the chaos by focusing on the smaller

---

This is a Profile of a recently elected member of the National Academy of Sciences to accompany the member's Inaugural Article on page 12299.

www.pnas.org/cgi/doi/10.1073/pnas.0505819102

units of protein structure, the motifs. She characterized how specific motifs within a structure related to sequence. Along with colleague Mike Sternberg, she published several articles examining how certain features of amino acids, especially chirality, could affect protein conformations (4–6). One study described the direction of the connecting turns in a β–α–β motif. Thornton found that these motifs almost always took right-handed turns (5). With a chuckle, she notes "that finding was really nice, because that rule is actually obeyed about 95% of the time."

In 1979, Thornton left Phillips' group and moved on to Birkbeck College (London) as an advanced fellow in the laboratory of Tom Blundell, a well known and respected structural biologist. She continued studying sequence–structure relationships and examined the molecular interactions that stabilize tertiary protein structure, such as salt bridges and disulfide bonds. Thornton stayed on at Birkbeck for over 11 years, first as a fellow and then as a lecturer. She primarily worked part-time so she could spend time at home with her family, balancing the chaos of proteins with the chaos of children.

Structuring Sequence Data

Structural biology had come a long way by the time Thornton began her first professorship in 1990 at University College London, in the Department of Biochemistry and Molecular Biology. Several hundred protein structures were now available for study, with more being solved each day, and advances in molecular biology provided a robust number of novel protein sequences with which to work. Although all the emerging data made some aspects of bioinformatics easier, the research became more complex because of the influx of massive amounts of sequence and structure data.

To organize and handle all these data, Thornton turned to her computer skills and helped design various computer programs and databases. She noticed that many proteins that appeared unrelated based on sequence had similar structures and/or functions. Although the sequence of amino acids of a protein may change drastically during evolution, Thornton observed that the structures adopted by these sequences were conserved (7). Along with Christine Orengo and David Jones, she discovered that only nine protein folds were known to recur in proteins having neither sequence nor functional similarity, and these folds dominated the database, representing more than 30% of all determined structures at that time.

Thornton and her group devised a clever method, termed "protein threading," to predict a protein's tertiary structure by using these protein folds (8). The technique involved threading a protein sequence onto the frameworks of known protein folds and finding the most energetically favorable conformation. In 1997, after years of work, Thornton also designed a new classification system for proteins based on structure instead of sequence, called CATH (Class, Architecture, Topology, Homologous Superfamily) (9).

As techniques improved, researchers also found that the structures for many proteins in the database contained inconsistencies or local errors. Because Thornton relied heavily on these older structures, she decided that an improved method to validate structural data was needed. In 1993, she and colleagues, including Roman Laskowski, designed a computer program, PROCHECK, that examines the stereochemistry of a protein's structure and compares its quality with structures of similar resolution (10). PROCHECK has since become one of Thornton's most notable applications and has been cited nearly 5,000 times in the literature.

Thornton and her colleagues at University College London later realized that their bioinformatic protein modeling programs could be highly useful in the pharmaceutical field. In 1998, she and colleagues helped start up the small drug discovery company Inpharmatica (London), which would utilize their software for drug design. "Our idea was to focus on using bioinformatics and chemoinformatics to improve our ability to identify targets for drug discovery," she says.

After Inpharmatica was set up, however, Thornton left the actual discovery to other researchers and settled into an advisory role with the company. "I'm an academic at heart," she says. "Drug discovery requires a lot of serendipity, and you're almost doing the impossible, whereas academic science is finding the things that are feasible and doable, and I much prefer that."

A Biologist Emerges

In 2001 Thornton took on the challenge of becoming Director of the European Bioinformatics Institute (EBI), which is part of the European Molecular Biology Laboratory (EMBL). The EBI looks after core molecular data in biology, providing the European center for many international database consortia, such as UniProt and wwPDB. Having based her career on protein structural data, Thornton appreciates its value and wishes to help ensure it receives the care and attention it deserves in terms of simple deposition, reliable curation and annotation, and easy access. "EBI combines the handling of these 'gold nuggets' of data with a thriving community of young bioinformatics research groups, who seek to unravel the mysteries the data hold," she says. "Combining knowledge derived from many different types of biological data will be the key to unraveling the molecular basis of life."

As Thornton explains, "For me, it's been a journey of being a physicist to being a biologist." Over the past several years, she has shifted her research to continue this trip into the life sciences. "The function is the critical thing about proteins," she says. "What they do and how they do it. I've started looking at how the structure affects the function, and how proteins interact with other proteins or small molecules."

"For me, it's been a journey of being a physicist to being a biologist."

In her Inaugural Article (1), Thornton presents a unique method to improve the ability to assign function to a protein based on sequence data. Currently, she notes, most functional prediction relies solely on comparing the target protein with homologous sequences in the database. She has found, however, that it is extremely difficult to assign function based on homologues that share less than 40% similarity, partially because many protein families evolve over time by changing a few key residues. "These families become diverse," she notes. "Enzymes in particular will conserve their chemistry but change the substrates with which they bind."

According to Thornton, over one-third of all proteins in the Research Collaboratory for Structural Bioinformatics (RCSB) database do not have any closely related sequences that meet the 40% threshold. In these cases, another approach is needed to prevent incorrect predictions from appearing in the protein databases. Thornton's computational approach, iCSA (Inpharmatica Catalytic Site Atlas), uses the catalytic residues of an enzyme, extracted from the literature (11), to assign related proteins as having similar or different functions when combined with biochemical data. Her Inaugural Article concentrates on a few benchmark experiments using enzymes: "Enzymes are the

best example since they require certain amino acids for catalysis. Therefore, catalytic sites are the most conserved throughout evolution."

Evolution is another term that is new to a physicist-turned-biologist like Thornton. "As a physicist, one isn't interested in historical aspects, since the laws don't change," she says. "In biology, however, what has happened in the past has a huge influence in what we see now. So understanding the evolution of molecules will help us understand how organisms have evolved."

In 2002, Thornton launched a side project tackling some of these evolutionary questions. Together with groups at University College London, she began looking into the functional genomics of aging. "We are looking at transcriptome data to find the relationship between caloric restriction, which makes flies and mice and worms live longer, and what happens in the body at the molecular level," explains Thornton. "We're trying to bring all that information together from these three animals, which requires using all the bioinformatics tools at our disposal." This endeavor may be daunting, but the goal certainly appears reachable—even for an academic such as Thornton.

Nick Zagorski, *Science Writer*

1. George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., MacArthur, M. W., Porter, C. T., Al-Lazikani, B., Thornton, J. M. & Swindells, M. B. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 12299–12304.
2. Thornton, J. M. & Bayley, P. M. (1976) *Biopolymers* **15,** 955–975.
3. Perkins, W. J., Piper, E. A. & Thornton. J. M. (1976) *Comput. Biol. Med.* **6,** 23–31.
4. Sternberg, M. J. E. & Thornton. J. M. (1976) *J. Mol. Biol.* **105,** 367–382.
5. Sternberg, M. J. E. & Thornton. J. M. (1976) *J. Mol. Biol.* **110,** 269–283.
6. Sternberg, M. J. E. & Thornton. J. M. (1977) *J. Mol. Biol.* **110,** 285–296.
7. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994) *Nature* **372,** 631–634.
8. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Nature* **358,** 86–89.
9. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997) *Structure (London)* **5,** 1093–1108.
10. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton. J. M. (1993) *J. Appl. Crystallogr.* **26,** 283–291.
11. Porter, C. T., Bartlett, G. J. & Thornton, J. M. (2004) *Nucleic Acids Res.* **32,** D129–D133.