the data." But this method is imperfect because the results are heavily influenced by the choice of bin size. So, Nelson's team turned to hidden Markov models.

"Hidden Markov models have a long and illustrious history in the study of single ion channels, but recently they have also increasingly been the method of choice when analyzing single-molecule biophysics experiments," Nelson says. Hidden Markov models help scientists make inferences about some unobservable data (e.g., DNA states) based on a set of observable and noisy data (e.g., bead movements). The algorithm estimates the unknown rates by finding the values that make the observed pattern of data the most likely.

"For a physicist, it's really beautiful to see the same ideas getting recycled in very different contexts," Nelson says. "But we had a technical challenge, we couldn't just take it off the shelf and use it because the classic set up wasn't quite applicable." Hidden Markov modeling assumes that the noise in the observable data is purely random. However, in tethered particle analysis, this assumption is violated: the position of the bead in one moment depends on the position of the bead the instant before. So, Nelson's team made a new model—called a diffusive hidden Markov model—that accounts for this dependency.

The resulting estimates of the rates of looping formation and breakdown were robust; their rate estimates did not change when they re-analyzed the data after removing every other datapoint.

"I think their approach seems very novel and sound, and it's clear that by doing this they can obtain more accurate information about DNA looping kinetics," says **Taekjip Ha, PhD**, associate professor of physics at the University of Illinois at Urbana-Champaign. Ha has done work using hidden Markov modeling for single-molecule fluorescence studies not involving tethered molecules.

—By Kristin Cobb, PhD

Parsing PubMed

Text-mining tools such as iHOP (Information Hyperlinked Over Proteins) are doing for biological literature what hyperlinks and search engines do for the Internet: organizing interconnected information in a fast, intuitive, searchable manner. And in January 2007, the service started to provide daily updates—extending the information network by about 2,000 new papers every day.

It is a huge challenge to parse the literature on an ongoing basis, with thousands of new papers per week

With genes and proteins acting as hyperlinks between sentences and abstracts, a large part of the PubMed knowledge base becomes a giant, navigable information network, says **Robert Hoffmann, PhD**, a postdoctoral fellow at Sloan-Kettering Institute who started the iHOP project while a researcher at the Protein Design Group at the National Center for Biotechnology (CNB) in Madrid, Spain. "The new version provides current information on even more genes and chemical compounds, covering 1,500 organisms ranging from human and chimpanzee to yeast and HIV," Hoffman says. He and his colleagues also extended iHOP's results to include drug interactions, and they've provided new ways to interact with the data—such as displaying "breaking news" found in papers from the past two years.

Freely available online since 2004, iHOP parses millions of PubMed documents and selectively grabs information specific to 80,000 different biological molecules. The program displays a list of relevant sentences snagged from the parsed documents, effectively summarizing the interactions and functions of a given protein or gene. The user can also browse statistical overviews of interaction partners and associated drugs, collect interesting sentences into a logbook, and create graphical representations of the results.

The computational machinery behind iHOP has continually evolved since the program's introduction, Hoffman says.

The most important enhancement this year—daily updating—was also the most technically demanding, requiring the daily processing of about 2,000 new publications. "It is a huge challenge to parse the literature on an ongoing basis, with thousands of new papers per week," says **Chris Sander, PhD**, of the Computational Biology Center at Memorial Sloan Kettering Cancer Center. "Robert and our team can now do this as the result of new software running on a multiprocessor machine that is better suited to processing large-scale text data."

The problem, Hoffmann says, is that most parallel computing pipelines (known as Message Passing Interface frameworks) are designed for repeated number crunching, not the sort of memory-intensive, semantic database processing that text mining requires. So Hoffmann developed his own computational pipeline capable of annotating millions of documents within a few hours on an 80-node cluster, making daily iHOP updates a reality. "We're now in a good position to make the next move toward annotations of full text sources, as well as the algorithmic exploration of gene networks," Hoffmann says.

Text-mining tools such as iHOP are great for focusing on pertinent key fragments in the literature, says **Russ Altman, MD, PhD**, chair of the Department of Bioengineering at Stanford University. "There is so much published that it's hard to keep track of all the relevant information, especially in journals that end up having unexpectedly relevant material," Altman says. "iHOP is an example of an approach that helps biologists filter lots of literature."

iHOP is freely accessible at http://www.ihop-net.org/.

—By Regina Nuzzo, PhD □