CELL BIOLOGY

Finding the trees in the forest

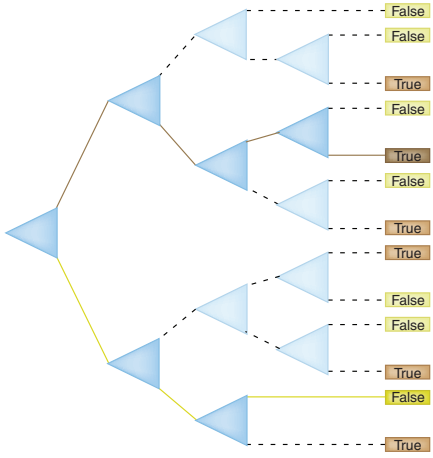
The integration of quantitative proteomics and analysis by machine learning yields a refined list of proteins involved in chromosome function.

Eight years ago Juri Rappsilber of the University of Edinburgh had a plan, beautiful in its simplicity. The then newly minted independent investigator decided to "plant his flag," as he describes it, in the chromatin field and to apply his knowledge in proteomics, acquired during years as a postdoc in Matthias Mann's laboratory, to identify proteins that bind chromosomes.

Three years into this work he met Bill Earnshaw, also from the University of Edinburgh, and both scientists felt that the combination of Rappsilber's proteomic expertise and of Earnshaw's biological background and chromosome purification skills would allow them to tackle the problem more efficiently. Rappsilber remembers their initial expectation: "We [planned to] just throw our two capabilities together and have a great paper in half a year." The expectation eventually materialized, but it took five more years.

An early disappointment came when the scientists looked at the initial results of a mass spectrometric analysis of all proteins that bound to the purified chromosomes. The list was too long, and known chromosomal proteins were not substantially enriched. Earnshaw describes the problem: "[After] nuclear envelope breakdown in mitosis, the chromosomes act like an ion-exchange resin that binds tightly large numbers of cytosolic proteins." The scientists realized that distinguishing proteins that functionally bind from opportunistic hitchhikers would need additional experiments.

Earnshaw, Rappsilber and their teams—led by Shinya Ohta and Jimi-Carlo Bukowski-Wills—decided to use stable-isotope labeling with amino acids in cell culture (SILAC) to monitor protein exchange. They hypothesized that proteins that are dynamically exchanged are more likely to constitute background. The researchers purified chromosomes from chicken cells and incubated them with mitotic cell lysate that included the heavy amino acids [^{13}C]Arg and [^{13}C]Lys. Then they ran the chromosome-associated proteins through a mass spectrometer. But still, this experiment did not give a clear-cut answer as to which proteins were bona fide functional binders.



Schematic of a random forest. Image was suggested by Jimi-Carlo Bukowski-Wills.

"We needed an approach that allowed the combination of multiple classifiers," says Earnshaw. They termed the approach multiclassifier combinatorial proteomics, or MCCP.

In addition to protein lists for the classifiers 'abundance', 'enrichment' and 'exchange', the researchers generated results with cell lines devoid of either condensin, a protein with a role in chromosome architecture, or Ska3, a recently identified kinetochore-associated protein with controversial function.

Then came the challenge of data analysis. Initially the teams plotted the data in three dimensions. Rappsilber recalls its limitations: "It's a very manual way that does not work beyond three classifiers." It also requires data points for all proteins in all experiments—something proteomics cannot deliver. The researchers settled on random forest analysis, a machine-learning approach that can deal with missing values. They trained decision trees on a set of proteins that they knew to be either cytosolic or chromosome bound, and then ran all the proteins they had found through the 'forest' of trees. This yielded a confidence score that indicated each protein's likelihood of being chromosomal.

Looking at the final lists, Earnshaw recalls being surprised at the number of new, high-confidence proteins. For the centromere-associated proteins alone, they predicted around 100 new candidates. To validate their strategy, the teams tested 50 proteins not previously characterized as chromosome binders in independent assays; 88% behaved as predicted.

Another pleasant surprise came when they looked at the effect of knocking out the condensin complex or Ska3. Removing either of these proteins also removed all their interaction partners from the chromosome. Thus the components of an insoluble structure could be deduced from what was missing in the knockout experiments—without having to isolate a large complex of proteins.

The final list of 4,000 likely functional, chromosome-binding proteins is a valuable resource for the community. One of many interesting findings is the candidates for bookmarking, the retention of transcription factors at promoters during mitosis.

But the scientists see value in their work beyond the resource. "I hope that people will see that this method is applicable to any complicated dataset that needs multiple classifiers to distinguish its members," says Earnshaw. Rappsilber sees it as a bit of an eye opener for the proteomics field. "You don't have to be able to plot data in two dimensions to get information out of it," he says. His final advice: "Do not fear creating complicated datasets."

Nicole Rusk

RESEARCH PAPERS

Ohta, S. *et al.* The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* **142**, 810–821 (2010).