Unravelling nature's networks

Nicholas A. M. Monk

Centre for Bioinformatics and Computational Biology, and Division of Genomic Medicine, University of Sheffield, Royal Hallamshire Hospital, Sheffield, S10 2JF, UK

Abstract

Dramatic progress has been made recently in determining the genetic and molecular composition of cells. This has prompted the development of new approaches to the challenge of understanding how basic cellular mechanisms are coordinated to produce the dazzling complexity of living systems. To face this challenge fully, it is critical not only to know what genes and proteins are expressed in cells, but also to understand the spatio-temporal dynamics of their networks of interactions. The sheer scale and complexity of cellular interaction networks necessitates a multi-disciplinary effort in which sophisticated experimental techniques are employed in combination with computational analysis and mathematical modelling. Such approaches are beginning to provide insight into basic structures and mechanisms, and promise to become critical to the post-genomic mission of understanding the cell as a complex dynamical system.

Determining network structure

The basic topology of an interaction network is determined by the nodes (e.g. the genes or proteins) and by the edges between these nodes (interactions between genes and/or proteins). In addition to this, it is often appropriate to label the edges to show directionality of interactions (e.g. a transcription factor binding to a promoter) and the nature of an interaction (e.g. activating or inhibitory). The nodes and edges constitute the network (or graph). The network of all interacting species within a cell is enormous, and a full determination of its structure remains beyond the reach of current experimental techniques due to technical and practical constraints. However, in practice it is often more appropriate to focus on smaller sub-networks that play a direct role in a particular setting or process (such as the network of interactions involving the transcription factor NF-κB in inflammation [1]). The structure of such sub-networks has traditionally been elucidated by intensive genetic and biochemical analysis. This results in reasonably detailed pictures of small fragments of the entire cellular network, while the structure of the bulk of the network remains unexplored.

Advances in high-throughput technologies now make it possible to obtain data about cells and tissues on genomic and proteomic scales. With information available on the expression and interactions of thousands of genes and proteins, the underlying structure of the complete cellular interaction network can begin to be explored systematically. The data can take a number of different forms, each with its own strengths and weaknesses. Examples include the phenotypic effect of 'knocking down' the activity of individual genes [2], relative mRNA expression levels [3,4], and physical interactions between proteins [5–9]. Each type of data can be interpreted in terms of interactions between genes and proteins, whether these are (possibly indirect) functional interactions or direct biochemical interactions.

One advantage of these high-throughput data for the purposes of constructing interaction networks is that they provide samples of the entire network, rather than being focused around a small number of nodes. These techniques do, however, have limitations. For example, identification of protein-protein

interactions by yeast two-hybrid and mass spectrometry of affinity-purified complexes is prone to significant false positives and negatives [10]. Furthermore, existing screens using these techniques provide limited coverage of all interactions, as illustrated by the surprisingly small overlap between different data sets [10]. Analysis of gene expression using DNA arrays is limited by the fact that relatively large amounts of RNA are required for the technique. This makes it difficult to perform studies on expression profiles in single cells or small populations of cells within a complex tissue [11]. An additional limitation is presented by the fact that only data at a single time point are produced. This can be overcome in some cell populations that exhibit synchronous dynamical behaviour, allowing time-course data sets to be obtained [12,13].

The impact of problems relating to data quality and coverage can be reduced by integrating distinct data sets. For example, combining data sets on protein-protein interactions has been demonstrated to increase data quality at the expense of coverage [10]. Different types of data sets may also be integrated, such as yeast two-hybrid protein-protein interaction data, DNA array gene expression data, and gene knock down data [14–16]. Such integration is important for network deduction, since techniques such as the yeast two-hybrid assay can give misleading information due to the fact that while two proteins may be capable of interacting, they may not be co-expressed in the cell type of interest. Clearly, however, some data are lost at each stage of integration, thus reducing network coverage.

The techniques available for the reconstruction of interaction networks from raw data depend on the type of data. Protein-protein interaction data specify a network of interactions directly, since edges in the network simply represent these interactions. Gene regulatory networks, however, are not uniquely defined by gene co-expression data, and network reconstruction can be achieved using a wide range of computational techniques [17]. A number of promising approaches use supervised learning on time-series data to deduce rules encoding the co-dependence of gene expression [18–20]. An alternative approach uses genome-wide location analysis to determine which genes have regulatory sequences to which each of a panel of transcription factors can bind, thus defining putative transcriptional networks [21].

Structural features of cellular networks

The properties of many different types of network, such as those based on social interactions, the internet and the world-wide web have been studied in detail. A striking finding that has emerged from these studies is that naturally occurring networks fall into a small number of classes, which can be defined using statistical measures of network structure [22]. Networks in these classes have structures (i.e. patterns of connectivity) that differ markedly to those found in simple randomly-connected networks [23], suggesting that their non-random structure reflects widespread underlying principles of organisation.

A prominent feature of biochemical networks is the fact that they appear to be scale-free, such that sub-networks of all sizes have the same statistical properties. A network is classified as scale-free if the distribution of degrees of its nodes (where the degree of a node is the number of edges to/from that node) follows a power law [24]. This property has been observed in a wide range of genomic and biochemical data sets [24–29]. A second widespread feature of naturally occurring networks is the 'small world' property, which encodes the fact that any two nodes in the network are connected by short paths that run along a small number of existing edges [22,24,30,31]. A characteristic feature of small world networks is that nodes tend to be found in local highly-connected clusters, with occasional edges linking distinct clusters [30].

The global statistical features of scale-free and small world networks are exhibited by a remarkable number of natural networks. Thus, while the occurrence of these features in biochemical networks may suggest that certain organisational principles underlie their functioning, their near universality makes it difficult to assess what they imply about the biochemistry of any particular case. Models of the evolution of scale-free networks have been proposed [27,32], but it is not clear to what extent these apply to cellular interaction networks. What is more, it is important to keep in mind that existing data sets represent only low-density samples of the entire network of cellular interactions. Low-density sampling of random networks that do not exhibit power law degree distributions can generate networks that do have such distributions, demonstrating that the statistical properties of networks obtained by sampling

may not give a true reflection of the properties of the underlying network [28]. Ultimately, additional classifiers of network structure are needed, based both on global and local network statistics. As an example of a local statistic, many well-documented regions of the yeast protein-protein interaction network exhibit a bipartite structure rather than 'small world' clustering [28]. How local structures relate to global statistics requires further investigation.

Organisational features of cellular networks

Global statistical measures such as node degree distribution and average clustering give a broad outline of network structure and suggest that biochemical networks are organised in a non-random fashion. More detailed studies are required to begin to determine the organisational principles underlying these networks. In attempting such an analysis, it is important to bear in mind that existing networks have evolved through natural selection acting on random mutations, and that any organisational principles must also have arisen by this route.

A strong candidate for a central organisational principle is modularity. Modularity implies that a network can be decomposed (formally) into functionally-separable sub-networks that exhibit a restricted set of dynamic behaviours corresponding to specific biological functions [33,34]. The existence of modules can be seen on many levels, examples being organelles and signal transduction pathways. While these examples show how modules can be identified through their physical localisation or known function, high-throughput data pose the challenge of identifying modules without recourse to known functional characteristics of a network. This has recently been achieved for a number of cellular networks using modified forms of cluster analysis that assign network nodes to modules [35–39]. In contrast to standard cluster analysis, in which nodes are assigned to a single cluster, modules overlap so that a node can be a member of multiple modules. These studies have revealed that a modular architecture can be discerned in bacterial gene expression [35], yeast gene expression and protein interactions [36–39], and metabolic networks in both prokaryotes and eukaryotes [40]. In the case of the yeast gene expression data, it is also possible to define the regulatory inputs to each module [37], allowing the higher order structure of the network in terms of its modules to be discerned.

The modularity of cellular networks appears, at first sight, to be incompatible with the evidence suggesting that these networks are scale-free, since modularity would be expected to impose a characteristic scale. One way of avoiding this conflict is to order modules in a hierarchical fashion, resulting in a modular network that preserves the power law degree distribution characteristic of scale-free networks [40]. Such a hierarchical structure has been observed in the metabolic networks of 43 distinct organisms, suggesting that it may represent a widespread organisational principle.

While modules correspond to functionally-separable units within a network, there is also evidence suggesting that gene regulatory networks contain conserved structural units, or motifs [41–43]. These motifs, which contain a handful of genes, are common to both bacteria and yeast. Examples include the feedforward loop and the bi-fan [41–43]. Interestingly, comparison of the sequences of the genes comprising specific examples of each class of motif suggests that they have not arisen by divergent evolution from common ancestral motifs, but that multiple motifs have arisen by convergent evolution. This in turn suggests that these common motifs represent elements of optimal circuit designs [43,44]. More generally, it has been suggested that modularity is intimately connected with both robustness and evolvability [34,45]. The validity of these claims in the context of cellular networks warrants further investigation.

Dynamics of cellular networks

Cells and tissues are complex dynamical systems. However, the techniques outlined above emphasise static features of cellular networks, such as their structure. To develop an understanding of the dynamical behaviour of all but the most trivial of networks it is necessary to complement experimental investigation with mathematical modelling [46]. Modelling is an essential tool not simply because of the extent of biochemical networks, but also due to nonlinearity (co-operativity), feedback loops, and layered combinatorial regulation of molecular interactions [47]. The

sheer number of distinct molecular interactions taking place at any time within a typical cell precludes detailed modelling; even if there were sufficient data to establish a model, and unlimited computer time for simulation, it would be impossible to deduce any principles underlying cellular dynamics. Instead, it is necessary to exploit the modularity of cellular processes and focus attention on restricted sub-networks. Within this restricted setting, the principal role of modelling is to elucidate the relationships between the structure and dynamical behaviour of modules.

Given a network topology, defined by experimental techniques, the first issue that has to be addressed is that of the modelling formalism to be used [48]. At a basic level, the rate of transcription of both prokaryotic and eukaryotic genes is a continuous variable with a stochastic component due to intrinsic and extrinsic noise [49,50]. Stochastic simulations that take this into account are perhaps the most realistic form of simulations, although they often incorporate other assumptions such as uniform spatial distributions within a cell. Furthermore, stochastic systems are often difficult to treat analytically. An alternative approach is to assume that the rates of molecular reactions are determined precisely by the (non-stochastic) concentrations of reactants. This results in systems of rate equations described by differential equations [46,48, 51–54]. Such systems can be very powerful, since a wide range of mathematical techniques are available with which they can be analysed to reveal the dependence of dynamical behaviours on features of the model. In addition to specifying the topology of the network to be modelled, it is necessary also to specify the values of the parameters that encode the properties of the interactions in the network [46]. This is not always possible, in which case simulations can be performed repeatedly using parameters drawn randomly from some biologically plausible distribution [55]. However, a range of techniques is now being developed to allow model parameters to be deduced using specific perturbations of cellular dynamics [46, 56–58].

Mathematical models have been used extensively to explore the dynamics of networks of moderate size [51–55]. Their validity can be tested by comparison with in vivo systems, where data of sufficient spatio-temporal resolution are available. In addition, it has recently become possible to design and construct small networks containing fluorescent reporters and to express these in prokaryotic cells [59–64]. These 'designer networks' allow model predictions and network behaviour to be compared directly, and have so far revealed good agreement between the two, both for stochastic and deterministic models. Nonetheless, care needs to be taken when specifying the mathematical form of models. For example, it is commonly assumed that molecules can exert their influence instantaneously. However, in reality processes such as transcription in eukaryotes involve substantial time delays; failure to incorporate these delays into models can make it impossible to capture observed network dynamics such as oscillatory gene expression without invoking extra network components [65,66].

A further complication arises from the fact that the functioning of intracellular networks can sometimes depend critically on the non-uniform distribution of their constituents [1]. These distributions may be static (many signal transduction pathways, for example, are organised by 'scaffold' proteins) or dependent on the activity of the network itself. A striking example of the latter case is provided by the network controlling chemotaxis in the slime mold *Dictyostelium*, which exhibits dynamic polarisation of signalling activities [67]. It is likely that the functioning of many networks involves significant spatial dynamics, and novel imaging techniques now allow these dynamics to be visualised [68–70].

Summary

Biological systems provide many striking illustrations of the emergence of robust system-level behaviour from underlying interactions that are potentially quite variable. Understanding the origins of robustness, and reasons for its failure, is a critical and challenging problem. The dynamic behaviours of cells and tissues depend on the operation of complex regulatory networks involving gene transcription and protein-protein interactions. Great strides are being made towards the goal of identifying and classifying the molecular components and overall structure of these networks, and their modular structure is being exploited in complementary experimental and modelling studies that provide detailed insight into the basic mechanisms underlying

network dynamics. Decomposing networks into constituent modules provides only a fragmented picture, however. To move towards a more complete picture of networks, it will be necessary to understand the ways in which modules can be integrated into larger networks so as to preserve their characteristic behaviour while allowing novel network-wide behaviours to emerge.

Acknowledgements

I gratefully acknowledge the support of the University of Sheffield (J.G. Graves Research Fellowship). Due to limitations of space, I apologise that I am unable to refer directly to all relevant primary research.

References

1. Dower, S.K. and Qwarnstrom, E.E. (2003) Biochem. Soc. Trans. this issue

2. Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapink, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003) Nature 421, 231–237

3. Baldi, P. and Hatfield, G.W. (2002) DNA Microarrays and Gene Expression, Cambridge University Press, Cambridge

4. Panda, S., Sato, T., Hampton, G.M. and Hogenesch, J.B. (2003) Trends Cell Biol. **13**, 151–156

5. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R.,
Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000) Nature **403**, 623–627

6. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) Proc. Natl. Acad. Sci. U.S.A **98**, 4569–4574

7. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002) Nature **415**, 180–183

8. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002) Nature **415**, 141–147

9. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlinmaier, S., Houfek, T., et al. (2001) Science **293**, 2101–2105

10. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Nature **417**, 399–403

11. Mills, J.C., Roth, K.A., Cagan, R.L. and Gordon, J.I. (2001) Nature Cell Biol. **3**, E175–E178

12. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O. , Botstein, D. and Futcher, B. Mol. Biol. Cell **9**, 3273–3297

13. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) Mol. Cell **2**, 65–73

14. Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F.C.P. (2002). Mol. Cell **9**, 1133–1143

15. Kemmeren, P. and Holstege, F.C.P. (2003) Biochem. Soc. Trans. this issue

16. Walhout, A.J.M., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K.C., Schetter, A.J., et al. (2002) Curr. Biol. **12**, 1952–1958

17. D'haeseleer, P. Liang, S., and Somogyi, R. (2000) Bioinformatics **16**, 707–726

18. Soinov, L.A., Krestyaninova, M.A. and Brazma, A. (2002) Genome Biol. **4**, R6.1–R6.10

19. Soinov, L.A. (2003) Biochem. Soc. Trans. this issue

20. Husmeier, D. (2003) Biochem. Soc. Trans. this issue

21. Lee, T.I, Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002) Science **298**, 799–804

22. Newman, M.E.J. (2003) SIAM Rev. **45**, 167–256

23. Erdös, P. and Rényi, A. (1960) Publ. Math. Inst. Hung. Acad. Sci. **5**, 17–61.

24. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A.-L. (2000) Nature **407**, 651–654.

25. Jeong, H., Mason, S., Barabási, A.-L. and Oltvai, Z. N. (2001) Nature **411**, 41–42.

26. Wagner, A. (2001) Mol. Biol. Evol. **18**, 1283–1292.

27. Pastor-Satorras, R., Smith, E. and Solé, R.V. (2003) J. Theor. Biol. **222**, 199–210

28. Thomas, A., Cannings, R., Monk, N.A.M., and Cannings, C. (2003) Biochem. Soc. Trans. this issue

29. Luscombe, N., Qian, J., Zhang, Z., Johnson, T. and Gerstein, M. (2002) Genome Biol. **3**, 0040.1–0040.7.

30. Watts, D.J. and Strogatz, S.H. (1998) Nature **393**, 440–442

31. Fell, D.A. and Wagner, A. (2000) Nature Biotech. **18**, 1121–1122

32. Barabási, A. L. and Albert, R. (1999) Science **286**, 509–512

33. Needham, J. (1933) Biol. Rev. **8**, 180–223

34. Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999). Nature **402**, C47–C52

35. Snel, B., Bork, P. and Huynen, M.A. (2002) Proc. Natl. Acad. Sci. U.S.A. **99**, 5890–5895

36. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Nature Genet. **31**, 370–377

37. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) Nature Genet. **34**, 166–176

38. Rives, A.W. and Galitski, T. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 1128–1133

39. Maslov, S. and Sneppen, K. (2002) Science **296**, 910–913

40. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A. L. (2002) Science **297**, 1551–1555

41. Shen-Or, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Nature Genet. **31**, 64–68

42. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Science **298**, 824–827

43. Conant, G.C. and Wagner, A. (2003) Nature Genet. **34**, 264–266

44. Csete, M.E. and Doyle, J.C. (2002) Science 295, 1664–1669

45. Hansen, T.F. (2003) BioSystems **69**, 83–94

46. Stark, J., Brewer, D., Barenco, M., Tomescu, D., Callard, R. and Hubank, M. (2003) Biochem. Soc. Trans. this issue

47. Weng, G., Bhalla, U.S. and Iyengar, R. (1999) Science **284**, 92–96.

48. Bower, J.M. and Bolouri, H., eds. (2001) Computational Modelling of Genetic and Biochemical Networks, MIT Press, Cambridge, Massachusetts

49. Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. (2002) Science **297**, 1183–1186

50. Blake, W.J., Kærn, M., Cantor, C.R. and Collins, J.J. (2003) Nature **422**, 633–637

51. Hasty, J., McMillen, D., Isaacs, F. and Collins, J.J. (2001) Nature Rev. Genet. **2**, 268–279

52. Tyson, J.J., Chen, K.C. and Novák, B. (2003) Curr. Opin. Cell Biol. **15**, 221–231

53. Novák, B. and Tyson, J.J. (2003) Biochem. Soc. Trans. this issue

54. Cho, K.-H. and Wolkenhauer, O. (2003) Biochem. Soc. Trans. this issue

55. von Dassow, G., Meir, E., Munro, E.M. and Odell, G.M. (2000) Nature **406**, 188–192

56. Ronen, M., Rosenberg, R., Shraiman, B.I. and Alon, U. (2002) Proc. Natl. Acad. Sci. U.S.A. **99**, 10555–10560

57. Gardner, T.S., di Bernardo, D., Lorenz, D and Collins, J.J. (2003) Science **301**, 102–105

58. Tegnér, J., Yeung, M.K.S., Hasty, J. and Collins, J.J. (2003) Proc. Natl. Acad. Sci. U.S.A. **100**, 5944–5949

59. Elowitz, M.B. and Leibler, S. (2000) Nature **403**, 335–338

60. Gardner, T.S., Cantor, C.R. and Collins, J.J. (2000) Nature **403**, 339–342

61. Becskei, A. and Serrano, L. (2000) Nature **405**, 590–593

62. Guet, C.C., Elowitz, M.B., Hsing, W. and Leibler, S. (2002) Science **296**, 1466–1470

63. Atkinson, M.R., Savageau, M.A., Myers, J.T. and Ninfa, A.J. (2003) Cell **113**, 597–607

64. Hasty, J., McMillen, D. and Collins, J.J. (2002) Nature **420**, 224–230

65. Monk, N.A.M. (2003) Curr. Biol. **13**, 1409–1413

66. Lewis, J. (2003) Curr. Biol. **13**, 1398–1408

67. Kriebel, P.W., Barr, V.A. and Parent, C.A. (2003) Cell **112**, 549–560

68. Miyawaki, A. (2003) Dev. Cell **4**, 295–305

69. Meyer, T. and Teruel, M.N. (2003) Trends Cell Biol. **13**, 101–106

70. Remy, I. and Michnick, S.W. (2001) Proc. Natl. Acad. Sci. U.S.A. **98**, 7678–7683