

Modularity map of the network of human cell differentiation

Viviane Galvão^{a,b}, José G. V. Miranda^b, Roberto F. S. Andrade^b, José S. Andrade Jr.^{c,d}, Lazaros K. Gallos^e, and Hernán A. Makse^{e,1}

^aDepartamento de Ciências Biológicas, Universidade Estadual de Feira de Santana, 44036-900, Feira de Santana, Bahia, Brazil; ^bInstituto de Física, Universidade Federal da Bahia, 40210-210, Salvador, Bahia, Brazil; ^cDepartamento de Física, Universidade Federal do Ceará, Campus do Pici, 60451-970, Fortaleza, Ceara, Brazil;

^dIFB, HIF E12, ETH Honggerberg, 8093 Zurich, Switzerland; and ^eLevich Institute and Physics Department, City College of New York, New York, NY 10031

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved February 4, 2010 (received for review December 22, 2009)

Cell differentiation in multicellular organisms is a complex process whose mechanism can be understood by a reductionist approach, in which the individual processes that control the generation of different cell types are identified. Alternatively, a large-scale approach in search of different organizational features of the growth stages promises to reveal its modular global structure with the goal of discovering previously unknown relations between cell types. Here, we sort and analyze a large set of scattered data to construct the network of human cell differentiation (NHCD) based on cell types (nodes) and differentiation steps (links) from the fertilized egg to a developed human. We discover a dynamical law of critical branching that reveals a self-similar regularity in the modular organization of the network, and allows us to observe the network at different scales. The emerging picture clearly identifies clusters of cell types following a hierarchical organization, ranging from sub-modules to super-modules of specialized tissues and organs on varying scales. This discovery will allow one to treat the development of a particular cell function in the context of the complex network of human development as a whole. Our results point to an integrated large-scale view of the network of cell types systematically revealing ties between previously unrelated domains in organ functions.

complex network | modular organization | self-similarity | stem cells

The cell differentiation process plays a crucial role in the prenatal development of multicellular organisms. Recent advances in the research on stem cell properties and embryonic development have uncovered several steps in the differentiation process (1–7). Single and multiple sequences of cell differentiation have been identified through *in vivo* observations of a particular embryo during early stages of development and through pathology studies of miscarriages during late stages of the process. While the identification of each cell differentiation step has been the subject of intense research, an integrated view of this complex process is still missing. Such a global view promises to reveal features associated with the large-scale modular organization of the cell types (5–12) with the purpose of discovering functional modules between cell types by using theoretical network analysis for community detection (9–11). In this letter, we take advantage of the current knowledge on the sequence of cell differentiation processes that is spread over a vast specialized literature (1–6, 13–27) (*SI Appendix*), to reveal and characterize the topological and dynamical features associated with the network of human cell differentiation (NHCD).

I. Results

We construct the NHCD by systematically gathering the scattered information on the evolution of each cell type present in the embryo and fetus from a predecessor with a higher degree of differentiation potential into a more specialized type. The process of cell differentiation is then mapped onto a complex network that consists of 873 nodes connected through 977 edges. The nodes in the network represent distinct cell types reported in the literature

(1–6, 13–27) and the edges represent the association between two cell types through a differentiation event.

The initial steps of the NHCD are shown in the *Inset* of Fig. 1, while the resulting network structure is shown in the main panel of Fig. 1 (*SI Appendix*). The fertilized egg is followed by the ball stage, and the formation of the primary germ cell layers. Currently, it is known that until the ball stage, cell division is symmetric and produces further totipotent stem cells (1). These cells then give rise to all the differentiated tissues of the organism as well as the extraembryonic tissues (placenta, umbilical cord, etc.). Moreover, in the course of the entire process of organism formation, there is a monotonic decrease in the differentiation potential (totipotent → pluripotent → multipotent → unipotent cells) accompanied with an increase in cell specialization.

Certain types of cells can be generated following more than one path from the fertilized egg. This process generates some closed loops of edges in the network. The NHCD comprises 529 branches of different lengths with each branch ending when the cell types do not undergo further differentiation. Note, however, that the most recent compilation of cell types in normal, healthy, human adults done in (6) reports only 407 cell types. Therefore, not all branch endpoints correspond to cell types in born humans. Thus, not all 873 cell types are present in a human being. Among those absent are the placenta cells that are generated from the fertilized egg during embryo development, as well as other somatic cell types that are important to control embryo and fetus development. The cell types that survive in a human are denoted by filled circles, while nonsurviving ones are indicated by empty circles. The complete collected data are listed in the *SI Appendix* and *Dataset S1*, including an enumeration of cells and links between the cell types, their time of appearance in days after fecundation (T_a), and the reference to the publications reporting each link. To the best of our knowledge, the structure identified here provides the most complete schematic diagram of the human differentiation process to date.

It is visually apparent from Fig. 1 that the NHCD has a prominent modular structure. The continuous differentiation of cells into more specialized functions naturally leads to the formation of dense isolated clusters in the NHCD. As a first approach to understand this modular structure, we cluster cell types in the network of Fig. 1 according to their known functions; different colors indicate 19 functional modules extracted from the literature (C1–C19) (*SI Appendix*). The largest communities were extracted from refs. 1–6, 13–27. There is, however, a certain degree of arbitrariness in this modular structure as the separation of the

Author contributions: V.G., J.M., R.F.S.A., J.S.A.J., L.K.G., and H.A.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. E-mail: hmakse@lev.cuny.cuny.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0914748107/DCSupplemental.

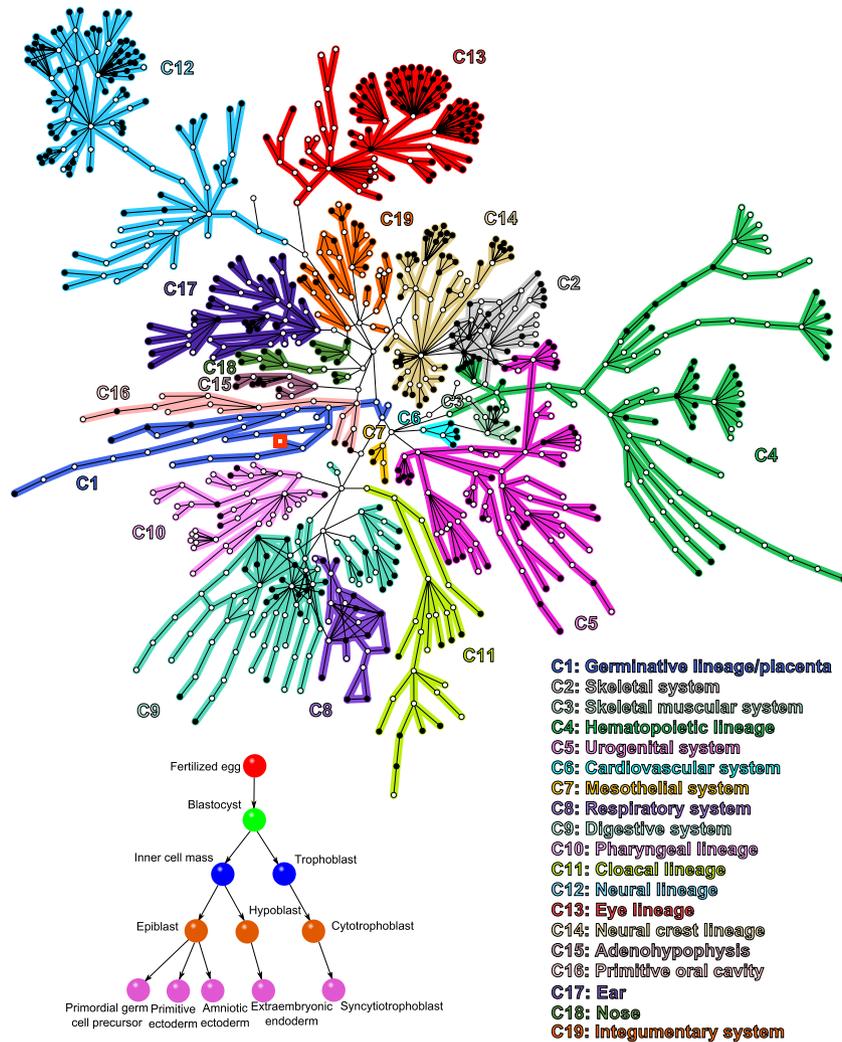


Fig. 1. Complex network representation of the human cell differentiation process. The first steps of the NHCD construction are shown in the *Inset* of this figure. These steps, known to also be present in the formation of the majority of multicellular organisms, include the first cleavage of a fertilized egg, that is, subsequently, followed by the ball stage and the formation of primary germ cell layers, namely, the ectoderm, mesoderm, and endoderm. The fertilized egg is a totipotent stem cell. The blastocyst, in turn, gives rise to both trophoblast and inner cell mass. These two cells further differentiate into other types of cells, and so on. Following the above process until the fetus is fully developed yields the complex network shown in this figure. Each node, plotted as a circle, corresponds to a cell type and the edges to a differentiation step. The entire network originates from the fertilized egg (denoted by a *Red Square*) and leads to the specialized cells of a developed human. *Filled Circles* correspond to nodes that survive at the end of the development process, whereas *Empty Circles* correspond to nonsurviving cell types. Nodes in communities of known functions from the literature are indicated by different colors, except for those cell types with no functional annotation (see *S1 Appendix* and *Dataset S1* for association to the known functions C1 to C19 extracted from the literature).

nodes into communities in our dataset is not unique. For instance, community C12, the neural lineage, could be divided into two subcommunities, representing the neural and the supporting (glial) cells (1–3, 5, 6, 1314). On the other hand, the neural system module could be merged with the eye system module (1, 5, 6, 13) on a larger scale because they have a common ancestral cell type. Therefore, a finer or coarser community structure can be extracted from the data when we look at the whole network at different scales of observation; a modified module-detection algorithm is needed to identify these communities in a systematic way.

Graph theoretical concepts allow us to unravel the scale dependence of the modular structure of the NHCD. Graph theory (11) defines the distance between two nodes (also called the chemical distance) as the number of links along the shortest path between the nodes in the network. We use this notion to propose a community detection algorithm that identifies modules of size ℓ composed of highly connected cell types. The algorithm finds the optimal tiling of the network with the smallest possible number of modules, N_B , of size ℓ (12) (each node is assigned to a module or

box and all nodes in a module are at distance smaller than ℓ). This process results in an optimization problem that can be solved by using the box-covering algorithm explained in Fig. 24, *Materials and Methods* Section III and reported in ref 28 as the Maximum Excluded Mass Burning algorithm (MEMB, the algorithm can be downloaded from http://lev.ccnycunyu.edu/~hmkase/soft_data.html). The requirement of minimal number of modules to cover the network (N_B) guarantees that the partition of the network is such that each module contains the largest possible number of nodes and links inside the module with the constraint that the modules cannot exceed size ℓ . This optimized tiling process gives rise to modules with the fewest number of links connecting to other modules implying that the degree of modularity, defined by (9–11, 29)

$$\mathcal{M}(\ell) \equiv \frac{1}{N_B} \sum_{i=1}^{N_B} \frac{L_i^{\text{in}}}{L_i^{\text{out}}}, \quad [1]$$

modules appear at all length-scales, i.e., modules are organized within larger modules in a self-similar way, so that the interconnections between those clusters repeat the basic modular character of the entire NHCD. Thus, the NHCD remains statistically invariant when observed at different scales. Varying the module size ℓ yields the scaling relation for the number of modules (Fig. 3B):

$$N_B(\ell) \sim \ell^{-d_B}, \quad [3]$$

where d_B represents the fractal dimension of the network (12). We find that the fractal character of the modules is established at the early stages, yielding $d_B \simeq 1.4$ as early as 30 d (Fig. 3B). As the network evolves, the fractal dimension increases slightly and finally reaches $d_B \simeq 1.9$.

The significance of Eq. 2 is that the modules need to be interpreted at a given length-scale. Figure 2B shows an example of such hierarchical organization (Fig. 2C and the *SI Appendix* show the full modular structure, while a list of detected modules appears in *Dataset S2*). Three types of communities of cell types are clearly identified in Fig. 2B as we change ℓ . (i) *The known functional modules*: The entire eye lineage (1, 5, 6, 13) is detected as a single module by the box-covering algorithm at $\ell = 11$, while the entire neural lineage (1–3, 5, 6, 13, 14) appears at $\ell = 15$.

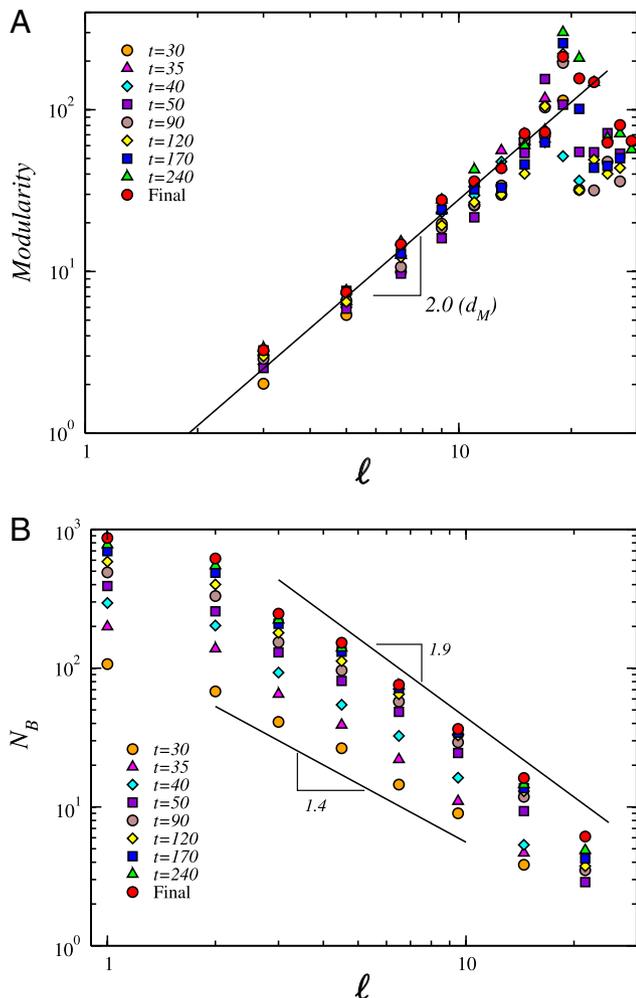


Fig. 3. Modular properties of the NHCD. (A) Degree of modularity of the network, $\mathcal{M}(\ell)$, at different times, t (indicated in the figure), as a function of the scale of observation, ℓ . (B) Number of boxes/modules, N_B , versus the size of the modules, ℓ , identified by the box-covering algorithm for different networks at time t .

Finer and coarser modules are identified by the algorithm. (ii) *Sub-modules*: At $\ell = 11$ the neural lineage is split into the main neural and the supporting glial cell modules, while for $\ell = 7$ sub-modules are identified in the eye system. (iii) *Super-modules*: When we increase the length to $\ell = 19$, the eye and neural system form a single super-module. Thus, each cell type is connected to other types according to which groups of nodes of all sizes self-organize following a single principle. This property allows us to renormalize the network (12) by replacing each detected module by a single supernode to identify the network of modules as shown in Fig. 2C. Following the evolution and interdependence of these super-modules, as seen in Fig. 2C, identifies families of cell types at varying scales. This modularity map is useful in proposing future research ties between previously unrelated domains in organ functions.

The dynamics leading to such a structure can be unraveled by the study of the NHCD as a growth process. The knowledge of the time of appearance of each cell type, T_a , makes it possible to follow the cumulative growth of the embryo and fetus in terms of the total number of cell types at time t , $N(t)$ as well as the number of cell types that eventually survive in the organism (Fig. 4A). As expected, surviving cells emerge in the later stages of the gestation period. However, the difference between the total and the surviving number of cell types indicates that generation of new types of nonsurviving cells takes place even during the final gestation months.

The increase of the network size, $N(t)$, is initially approximately exponential and after $t_* = 40$ days changes into a slower growth (Fig. 4A). Only a small percentage of the nodes grow within a given time interval so that the network activity is focused in a small number of them at a given time. The number of nodes that differentiate at a given time are shown in Fig. 4B. We observe an activity that increases monotonically up to around $t_* = 40$ days and then drops to lower values. The crossover time $t_* = 40$ days observed in Fig. 4A and B separates two regimes of growth and approximately corresponds to the time below which most of the cells have a plastic characteristic (i.e., the capability to differentiate) and above which they start to become functional (1). Interestingly, the two regimes observed in $N(t)$ merge into a single universal functional curve when we replot $N(\ell_{N1})$ as a function of the chemical distance to the fertilized egg, ℓ_{N1} (Fig. 4C). This result suggests that the topological distance in the network ℓ_{N1} is the natural variable to characterize the growth process in a universal form rather than the time. The dynamic of $N(\ell)$ follows a typical logistic (Verhulst) process of population growth where the rate of growth is restricted by environmental limitations:

$$\frac{dN}{d\ell} = rN \left[1 - \frac{N}{N_f} \right], \quad [4]$$

with solution,

$$N(\ell) = N_f \frac{\exp(r\ell)}{N_f + (\exp(r\ell) - 1)} \quad [5]$$

(see the fitting in Fig. 4C), where N_f is the final number of cell types and $r = 0.65$ is the growth rate of cell types.

Analysis of the network connectivity reveals that the average number of links per node in the final stages of the entire NHCD is $\langle k \rangle = 2.24$ (Fig. 4D). Even though $\langle k \rangle \approx 2$, there is a broad degree distribution [scale-free (11), $P(k) \sim k^{-\gamma}$, $\gamma \simeq 3.0$, and *SI Appendix*]. This implies that there is always a small number of crucial cell types that differentiate much more than the others, a fact that agrees with evidence on the existence of a few cells with large plasticity potential. As this potential is rapidly lost after 40 d, cell types change their development ability in favor of the organism life maintenance.

between all their nodes, at a given length-scale. Different values of the box diameter ℓ yield boxes of different size. These boxes are identified as modules that at a smaller scale, ℓ , may be separated but merge into larger entities as we increase ℓ .

In this work we implement the Maximum Excluded Mass Burning (MEMB) algorithm from (28) for box covering. The algorithm uses the basic idea of box optimization, where we require that each box should cover the maximum possible number of nodes, and works as follows: For a given ℓ , we first locate the optimal "central" nodes that will act as the origins for the boxes. This is done by first calculating the number of nodes (called the mass) within a diameter ℓ from each node. The node that yields the largest mass is marked as a center. Then we mark all the nodes in the box of this center node as "tagged." We repeat the process of calculating the mass of the boxes starting from all noncenter nodes, and we identify a second center according to the largest remaining mass, while nodes in the corresponding box are tagged, and so on. When all nodes are either centers or tagged we have identified the minimum number of centers that can cover the network at the given ℓ value. Starting from these centers as box origins, we then simultaneously burn the boxes from each origin until the entire network is covered, i.e., each node is assigned to one box [we call this process burning because it is similar to burning algorithms developed to investigate clustering statistics in percolation theory (11)]. In Fig. 2A we show how box covering works for a simple network at two different ℓ values.

This algorithm is driven by the proximity between nodes and the maximization of the mass associated with each module center (12, 28). Thus, it detects boxes that maximize modularity (Eq. 1). In the case of MEMB, we have the additional benefit of detecting modules at different scales so that we can study the hierarchical character of modularity, i.e., modules of modules, and

we can detect whether modularity is a feature of the network that remains scale-invariant.

The fractal dimension d_B of a complex network is an exponent that determines how the mass (equivalently: the number of nodes) around any given node scales with the length that in networks corresponds to the shortest distance between two nodes. To numerically measure this exponent we optimally cover the network with boxes using the MEMB algorithm. A box is a set of nodes where all distances ℓ_{ij} between any two nodes i and j in this set are smaller than a given value of ℓ , the box size. Although there is a large number of coverings, for every value of ℓ we want to find the one that gives the smallest possible number of boxes, $N_B(\ell)$. Varying ℓ then yields the scaling relation (Eq. 3). A finite fractal dimension reveals fundamental organizational principles of the underlying network, namely a self-similar structural character, where the network is built in a similar way even though we observe it at different length-scales. The boxes that are identified through this process correspond to the modules at varying scales.

ACKNOWLEDGMENTS. We thank J. Bruijć, B. Dubin-Thaler, H.D. Rozenfeld, S. Havlin, T. Rattei, A.M. Andrade, J.M.M. Andrade, and M. Sigman for valuable discussions. This work was supported by National Science Foundation Grants SES-0624116 and EF-0827508. We thank the financial support of the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Fundação de Amparo à Pesquisa do Estado da Bahia, and Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico, the National Institute of Science and Technology for Complex Systems, and the European Seventh Framework Programme Neuronano project (NMP4-SL-2008-214547).

- Sell S (2004) *Stem Cells Handbook* (Humana Press, Totowa, NJ).
- Kirschstein R, Skirboll LR (2001) *Stem Cells: Scientific Progress and Future Research Directions* (NIH, Bethesda).
- Freitas RA, Jr (1999) *Nanomedicine, Volume I: Basic Capabilities* (Landes Bioscience, Georgetown, Texas).
- Alberts B, et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York), 1Vth Ed.
- Sadler TW (2004) *Langman's Medical Embryology* (Lippincott Williams and Wilkins, Baltimore), 1Xth Ed.
- Vickaryous MK, Hall BK (2006) Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev* 81:425–455.
- Valentine JW (2003) *Keywords and Concepts in Evolutionary Developmental Biology*, eds BK Hall and WM Olson (Harvard Univ Press, Cambridge), pp 35–53.
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297:1551–1555.
- Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113.
- Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. *Nature* 433:895–900.
- Caldarelli G, Vespignani A, eds. (2007) *Large scale structure and dynamics of complex networks* (World Scientific, Singapore).
- Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. *Nature* 433:392–395.
- Paxinos G, Mai JK (2004) *The Human Nervous System* (Elsevier Academic, San Diego), 2nd Ed.
- Temple S (2001) The development of neural stem cells. *Nature* 414:112–117.
- Bianco P, Riminucci M, Gronthos S, Robey PG (2001) Bone marrow stromal stem cells: Nature, biology, and potential applications. *Stem Cells* 19:180–192.
- Chen JCJ, Goldhamer DJ (2003) Skeletal muscle stem cells. *Reprod Biol Endocrin* 1:101.
- Janeyway CA, Travers P, Walport M, Shlomchik M (2001) *Immunobiology: The Immune System in Health and Disease* (Garland Science, New York), 5th Ed.
- Anglani F, et al. (2004) In search of adult renal stem cells. *J Cell Mol Med* 8:474–487.
- Horster MF, Braun GS, Huber SM (1999) Embryonic renal epithelial: induction, nephrogenesis, and cell differentiation. *Physiol Rev* 79:1157–1191.
- Herrick SE, Mutsaers SE (2004) Mesothelial progenitor cells and their potential in tissue engineering. *Int J Biochem Cell B* 36:621–642.
- Otto WR (2002) Lung epithelial stem cells. *J Pathol* 197:527–535.
- Jessen KR, Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nat Rev Neurosci* 6:671–682.
- Nakashima M, Reddi AH (2003) The application of bone morphogenetic proteins to dental tissue engineering. *Nat Biotechnol* 21:1025–1032.
- Santagati F, Rijli FM (2003) Cranial neural crest and the building of the vertebrate head. *Nat Rev Neurosci* 4:806–818.
- Savage JJ, Yaden BC, Kiratipranon P, Rhodes SJ (2003) Transcriptional control during mammalian anterior pituitary development. *Gene* 319:1–19.
- Forge A, Wright T (2002) The molecular architecture of the inner ear. *Brit Med Bull* 63:5–24.
- Panteleyev A, Jahoda CAB, Christiano AM (2001) Hair follicle predetermination. *J Cell Sci* 114:3419–3431.
- Song C, Gallos LK, Havlin S, Makse HA (2007) How to calculate the fractal dimension in complex networks. *J Stat Mech* P03006.
- Gallos LK, Song C, Havlin S, Makse HA (2007) Scaling Theory of transport in complex biological networks. *Proc Natl Acad Sci USA* 104:7746–7751.
- Athreya KB, Ney PE (2004) *Branching processes* (Dover, New York).