Journal of Biological Research-Thessaloniki **18**: 353 – 358, 2012
J. Biol. Res.-Thessalon. is available online at http://www.jbr.gr
Indexed in: WoS (Web of Science, ISI Thomson), SCOPUS, CAS (Chemical Abstracts Service) and DOAJ (Directory of Open Access Journals)

The quest for an upper limit in systems biology

Christos A. OUZOUNIS[1,2,3*], Vasilis J. PROMPONAS[1] and Ioannis ILIOPOULOS[4]

[1] *Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, PO Box 20537, CY-1678 Nicosia, Cyprus*

[2] *Institute of Agrobiotechnology, Centre for Research & Technology Hellas (CERTH), GR-57001 Thessaloniki, Greece*

[3] *Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M5S 3E1, Canada*

[4] *Division of Medical Sciences, Medical School, University of Crete, GR-71110 Heraklion, Greece*

Received: 27 May 2012 Accepted: 29 May 2012

We discuss order-of-magnitude formulations for molecular systems biology, in order to derive estimates for the upper limit of the number of unique gene families and their ensuing potential interactions. A useful equation in the field of the Search for ExtraTerrestrial Intelligence (SETI), known as the Drake equation, can be mapped precisely to the problem of estimating the total number of unique gene families. Despite the fact that the parameter values cannot be accurately estimated at present, this semantic mapping provides a basis upon which a novel research agenda might be established, delimiting the scope of our technological capabilities in systems biology and appreciating the complexity of our scientific aspirations.

Key words: systems biology, gene families, Drake equation, protein interactions.

INTRODUCTION - BACKGROUND

Systems, Biology and Systems Biology

In recent years, there has been a rebirth of systems approaches to biological research, collectively known as "systems biology" (Ideker *et al*., 2001). Despite certain difficulties associated with the definition of this field (Cowley, 2004) or process (Naylor & Cavanagh, 2004), the goals of systems biology are highly ambitious: to represent, perturb and manipulate biological systems so that the functional roles of the individual components can be unveiled and ultimately understood in their entire context (Ideker *et al*., 2001). There has been a general agreement and a precipitous realization that the function of molecular components in cells is context-dependent (Burgess, 2004), and methods that specifically address this issue have been developed, for example the computational detection of functional modules in protein interaction networks (Spirin & Mirny, 2003; Pereira-Leal *et al*., 2004).

Where opinions differ is the limit of this wider context. Typically, most of systems biology has initially focused on molecular systems, but nothing in the original definition precludes the expansion of that scope towards organisms, populations, species or ecosystems: "A growing wave of biological research aims at systems - from networks of proteins in signal transduction pathways to communities of species" (May, 2005). In fact, it could be argued that the roots of systems biology arise from the pioneering work of von Neumann, Wiener, von Bertalanffy, Rosen and others (Cornish-Bowden & Cardenas, 2005). The earliest reference to a systems approach appears to be the definitive work of Lotka on physical biology, which dealt with species equilibria and ecological modeling on a grand scale, among other things (Lotka, 1925).

[*] Corresponding author: tel.: +30 2310 498473, e-mail: ouzounis@certh.gr

RESULTS AND DISCUSSION

Upper limits and orders of magnitude

In contemporary systems biology, one of the important parameters at the cellular level is the number of potential pairwise interactions for molecular components (e.g. genes or proteins). The number of interactions appears to be large, but an upper limit is hardly ever mentioned in the published record, with some notable exceptions (Grigoriev, 2003). Beyond the cellular level, there are very few accurate estimates, for instance the number of cells in the human body estimated at 10^{12} (these estimates vary, but we accept a widely cited figure here) or the number of bacterial cells within the human body at 10^{14} cells (Todar, 2008). Curiously, biologists are not accustomed to large numbers (a cursory Google™ search for "trillions" in "biology" yields 4 M entries, "quadrillions" 1.5 M entries and "quintillions" a mere 0.5 M entries - these numbers refer mostly to computing. For biological entities, other metrics are, of course, available). Yet, the largest number first cited anywhere in science more than three centuries ago was allegedly by Robert Hooke in 1665, who estimated the number of little chambers ("cells") for a square inch of cork at 1259712000 (Bryson, 2003). He defined these little chambers as "cells" because they reminded him the monks' quarters (Bryson, 2003). This is indeed an eerie coincidence with the subject matter of modern systems biology.

Order-of-magnitude formulations are usually confined in chemistry for molecules or astronomy for stars or galaxies, hence the term "astronomical", for truly enormous scales. Systems biologists need to take into account these scales, in order to delimit the domain of discourse. For example, astronomers estimate that the number of atoms in the visible universe stands at 10^{80} (answers.com, 2005) - for an explanation on how this number is calculated using density, volume and atomic numbers, see nso.edu (2001). What would be an equivalent question in systems biology? One such example can be the number of unique genes or gene families in our biosphere, reminiscent of the bet placed for the number of genes in the human genome (Rabinowicz *et al*., 2000; Goodman, 2003) but on another, truly planetary scale altogether.

We define a unique gene as a member of a gene family with a unique nucleotide sequence (e.g. a gene containing point mutations), and a unique gene family as a group of genes with no detectable similarity - thus implying homology - to other gene families. A unique gene family might contain multiple copies of unique genes across (or even within) species or populations. A unique gene without homology to any other gene might constitute a unique gene family.

From extraterrestrial civilizations to terrestrial gene numbers

How would one go about estimating the total number of gene families, thus setting an upper limit for the total number of possible gene or protein interactions (within or between species)? Previous work suggests that the number of newly discovered protein families has not saturated (Kunin *et al*., 2003), depending on the detection strategies (Kunin *et al*., 2005) and differential coverage by various databases (Ouzounis *et al*., 2003). Indeed, there are certain indications that gene family evolution is self-accelerating, as judged by sophisticated birth-and-death modeling of this process (Karev *et al*., 2004). These facts have been confirmed in a number of independent studies more recently (Peregrin-Alvarez & Parkinson, 2007; Tautz & Domazet-Loso, 2011) as well as experimental (Yooseph *et al*., 2007) or computational (Sammut *et al*., 2008) evidence. Thus, we need to seek a formulation that captures this diversity and explore the limitations of parameter estimation within this formulation.

Again, a source of inspiration comes from astronomy and the field of the Search for ExtraTerrestrial Intelligence (SETI). The Drake equation (Drake, 2004), first proposed to estimate the number of extraterrestrial civilizations in our Milky Way, has inspired not only the narrow field of SETI, but also other areas of science, such as linguistics - see for example Hook (2004). This equation, composed of seven terms, is stated as follows (Fig. 1):

$$N = R_* \times f_P \times n_e \times f_l \times f_i \times f_c \times L \qquad \text{[eq. 1]}$$

where N is the number of technological civilizations in the Milky Way, R_* the rate of formation of stars which could have planets with intelligent life, f_p the fraction of those stars with planets, n_e the average number of those planets per star that could sustain life, f_l the fraction of those planets with life, f_i the fraction of those planets with intelligent life, f_c the fraction of those planets with civilizations and L the length of time those civilizations can utilize technology before they disappear (for a Drake equation calculator, see http://aftercontact.org/2010/11/online-drake-equation-calculator-try-it-yourself/).

Importantly, the product ($R_* \times L$) is sometimes expressed as $\left(\frac{n_* \times L}{t_0}\right)$, where n_* is the current number of stars in the Milky Way and t_0 is the age of our own stellar system (Cirkovic, 2004). Although most of these parameters can only be approximated at best (wikipedia.org, 2012), this celebrated equation can in principle provide an order-of-magnitude estimate for the number of technological civilizations in our galaxy.

If we strictly maintain the structure for the Drake equation [eq. 1], we will need a direct semantic mapping of the astronomical terms to biological terms (Fig. 1). Thus, the equation remains as follows:

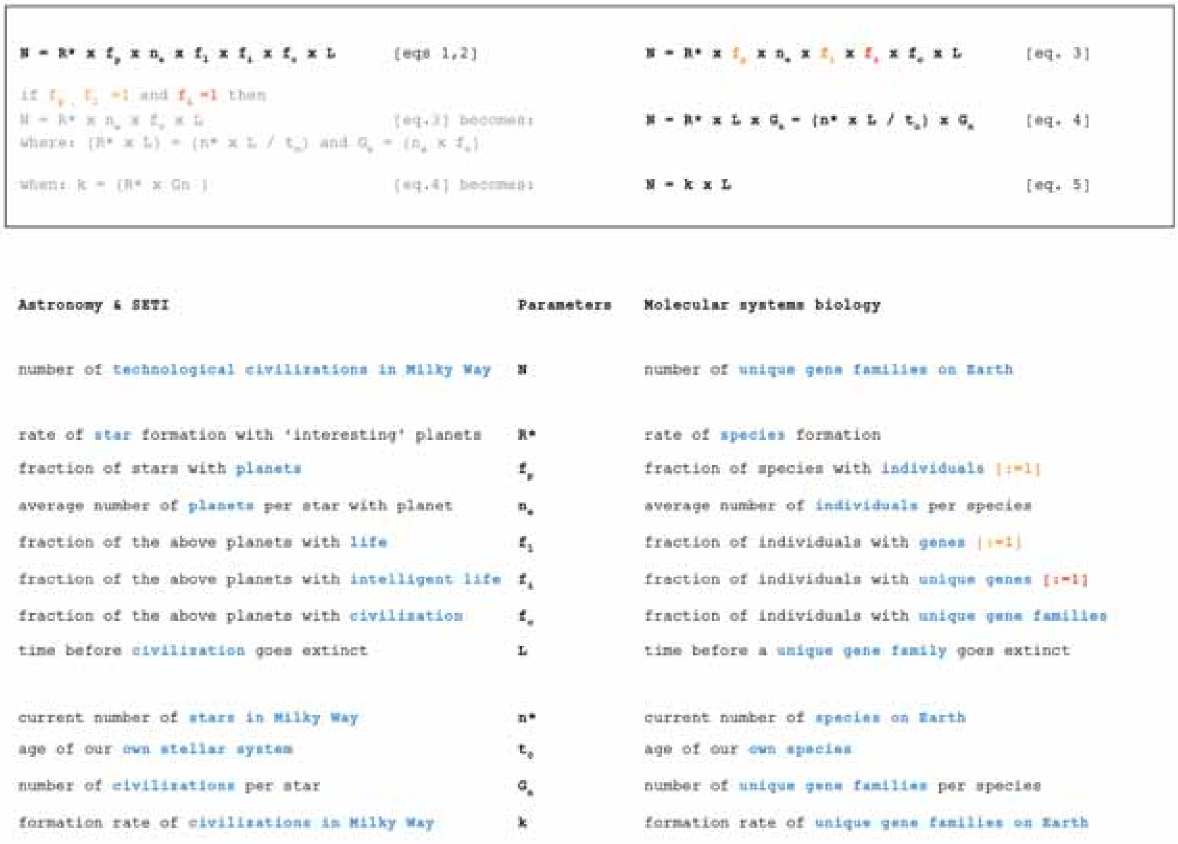
$$N = R_* \times f_P \times n_e \times f_l \times f_i \times f_c \times L \qquad \text{[eq. 2]}$$

where N is the number of unique gene families on Earth (or any biosphere), R_* the species formation rate, f_p the fraction of those species with individuals [:=1], n_e the average number of those individuals per species that are alive, f_l the fraction of those individuals with genes [:=1], f_i the fraction of those individuals with unique genes, f_c the fraction of those individuals with unique gene families and L the length of time those unique gene families exist before they go extinct (Fig. 1). Cells per individual are not taken into account in our version of the Drake equation (this would be the equivalent of living creatures per planet with life), so that the total number of unique genes (planets with intelligent life) is calculated on a per individual (planet), and not on a per cell (*cf* creatures per planet), basis.

Thus, the Drake equation for the number of unique gene families [eq. 2] is simplified to:

$$N = R_* \times n_e \times f_i \times f_c \times L \qquad \text{[eq. 3]}$$

which means that the number of unique gene families on Earth is a product of the unspecified parameter



$N = R^* \times f_p \times n_e \times f_l \times f_i \times f_c \times L$ [eqs 1,2]

$N = R^* \times f_p \times n_e \times f_l \times f_i \times f_c \times L$ [eq. 3]

if f_p , f_l =1 and f_i =1 then

$N = R^* \times n_e \times f_c \times L$ [eq.3] becomes:

$N = R^* \times L \times G_n = (n^* \times L / t_0) \times G_n$ [eq. 4]

where: $(R^* \times L) = (n^* \times L / t_0)$ and $G_n = (n_e \times f_c)$

when: $k = (R^* \times Gn)$ [eq.4] becomes:

$N = k \times L$ [eq. 5]

Astronomy & SETI	Parameters	Molecular systems biology
number of technological civilizations in Milky Way	N	number of unique gene families on Earth
rate of star formation with 'interesting' planets	R*	rate of species formation
fraction of stars with planets	f_p	fraction of species with individuals [:=1]
average number of planets per star with planet	n_e	average number of individuals per species
fraction of the above planets with life	f_l	fraction of individuals with genes [:=1]
fraction of the above planets with intelligent life	f_i	fraction of individuals with unique genes [:=1]
fraction of the above planets with civilization	f_c	fraction of individuals with unique gene families
time before civilization goes extinct	L	time before a unique gene family goes extinct
current number of stars in Milky Way	n*	current number of species on Earth
age of our own stellar system	t_0	age of our own species
number of civilizations per star	G_n	number of unique gene families per species
formation rate of civilizations in Milky Way	k	formation rate of unique gene families on Earth

FIG. 1. A schematic representation of the Drake equation for astronomy and the search for extraterrestrial intelligence (SETI) *versus* molecular systems biology and the quest for the number of unique gene families and ensuing interactions. The semantic mapping of terms across the two fields is highlighted. Equations in the text are numbered; the three estimates set to unity are shown in either orange (certain estimates) or red (uncertain estimate).

product ($R_* \times L$), corresponding to the effective number of living species, times individuals per species, times the two fractions of individuals per species with unique genes and families.

As already mentioned, the product ($R_* \times L$) can also be expressed as $\left(\frac{n_* \times L}{t_0}\right)$ where n_* is the current number of species and t_0 is the age of our own species. These values can be approximately set as follows, always on an order-of-magnitude scale: $n_* = 10^7$ and $t_0 = 10^6$, at least for macroscopic organisms from the geological record (AAAS, 2005), or simply 10 species yr^{-1} (over geological timescales). For the human species, these values are very different and can be set at $n_* = 1$ and $t_0 \approx 1$ Myr (or 10^6 yr), yielding an estimated value for $R_* = 10^{-6}$. Note that we do not have an estimate for L, yet, in this context.

To verify the logical validity of this argument, we can attempt to derive a smaller-scale estimate for the human species, solving for L [eq. 3]: with $R_* = 10^{-6}$, $n_e = 6 \times 10^9$ (human population), $f_i = 10^{-3}$ (frequency of single nucleotide polymorphisms, or SNPs), $f_c = 10^{-5}$ (frequency of large, rarer polymorphisms), and, for argument's sake, N set to 60, i.e. the number of unique gene families in the human genome - this number is reportedly 20 (Demuth *et al.*, 2006). In other words, the human genome might contain a number of unique genes (at higher frequency, corresponding to population variants) and a number of unique gene families (at lower frequency, corresponding to rare polymorphisms), according to our definitions above. In this case, $L = 10^6$, for humans, closely enough to an ultimate lifetime of the species. Not surprisingly, the effective species number ($R_* \times L$) in this calculation is equal to one. This relatively high estimate for L should be considered valid, given that it is taken to represent unique gene families with polymorphisms in the entire population, where individuals (and not species) are considered as the vehicles of novel gene families, few of which probably survive in the population background.

Of course, for other species, the hardest numbers to estimate are probably n_e and L, but the fractions/contributions can also be calculated on a per species basis only. If we set $f_i = 1$ (all individuals contain unique genes, possibly a reasonable assumption for the vast numbers of rapidly evolving species being the majority, e.g. bacterial populations), and merge the product of two other variables to G_n (i.e. one individual per species implying a clonal population, where contributions of unique families come from entire species, or in other words the "individuals" factor is taken out of the entire equation), then $G_n = n_e \times f_c$, where G_n is the number of unique gene families per species (Fig. 1). Thus:

$$N = (R_* \times L) \times G_n = \left(\frac{n_* \times L}{t_0}\right) \times G_n \qquad \text{[eq. 4]}$$

or the number of unique gene families is the number of unique gene families per species *times* the effective number of living species. Variables n_*, t_0 and G_n can further be represented as a single constant $k = R_* \times G_n \left(\frac{n_*}{t_0}\right) \times G_n$, signifying the formation rate of unique gene families. The values for n_*, t_0 and G_n can be estimated for our biosphere, as follows: $n_* = 10^7$, $t_0 = 10^6$, as above (AAAS, 2005), and, say, $G_n = 10^2$. These values represent the current number of known species (~10^7 species), the age of our own species (~10^6 yr), while the number of unique gene families per species is set arbitrarily to 100 (Demuth *et al.*, 2006). Parameter *k* might be estimated at $k \approx 1000$, according to the previous numbers. Ultimately (Fig. 1), our estimate for N [eq. 4] is reduced to:

$$N = k \times L \qquad \text{[eq. 5]}$$

where $N = 1000 \times L$. This is a significant, original result and a direct consequence of the Drake formula: it means that the number of unique gene families, under our previous assumptions, is the formation rate of unique gene families *times* their lifetime in any species before they are lost during evolution.

Previously, L was set at $L = 10^6$, but in this case, L might correspond to a much higher value, as the lifetime of a unique gene/protein family across species (not individuals). Even with $L = 10^6$, a minimum estimate for N is 10^9, or one billion unique protein families. Conversely, this equation [eq. 5] can be solved for any variable, assuming that N will be known. For instance, if we assume that $N = 10^3$, as previously proposed (Chothia, 1992), and we solve for:

$$G_n = \left(\frac{N \times t_0}{n_* \times L}\right) \qquad \text{[eq. 6]}$$

then $G_n = \left(\frac{10^3 \times 10^6}{10^7 \times 10^6}\right) = 10^{-4}$, implying that one out of every 10,000 species contributes a unique gene family, possibly a low estimate, given our previous discussion and the breadth of planetary biodiversity.

CONCLUSIONS

Future prospects

From this viewpoint, a possible research agenda for systems biology can be the inference of accurate estimates for L, because ultimately - as we argued above [eqs 2-6], where *k* is 1000 and according to this crude estimate for the variables n_*, t_0 and G_n - this most interesting value for parameter N will crucially depend on L, the average lifetime of a gene family (Karev *et al.*, 2004).

The significance of the Drake equation for the number of unique gene families is not so much in the accuracy of such an estimate - virtually an impossible task - but the realization of the immensity of the scope for an all-encompassing systems biology. If, for example, the number N is found to be of the order of 10^9, then the number of possible pairwise gene/protein cross-species interactions can be truly astronomical, at 10^{18}. The specificity of molecular recognition implies that the actual number of these interactions is much lower, given a vast scale of constraints, from stereochemistry to species barriers. Consequently, this unknown molecular landscape of our biosphere sets a valid and ambitious goal for biological science as a whole, from molecular biology to ecology. Molecular biologists rarely look beyond a species boundary, possibly due to experimental limitations; ecologists use molecular techniques as a toolkit for species or population markers, but do not necessarily investigate functional roles of molecules and their interactions on a large scale. Yet, the numbers we are faced with can be vast, including multi-species interactions as revealed by the human microbiome (Yang *et al.*, 2009).

The reasoning outlined here should in principle provide a unified view of gene families and ultimately contribute towards the realization that this number is large, yet finite. Molecular systems biology, which has limited itself to cells, might one day expand to encompass our entire biosphere (Veizer, 1988). Our present analysis should not be taken as a challenge to the current research agenda, but as a first step towards an accurate estimate for the upper limit of a planetary-scale molecular systems biology, the number of genes and their interactions.

ACKNOWLEDGEMENTS

This analysis was inspired by Isaac Asimov's book "Extraterrestrial civilizations" (Pan Books, 1981, London). We would like to thank Professor Frank Drake (SETI Institute, Director of the Center for the Study of Life in the Universe), for his encouragement and invaluable comments. We thank Nikos Darzentas, Leon Goldovsky, Pierre Maziére and Sophia Tsoka (all at the European Bioinformatics Institute, Cambridge UK) for their contributions of earlier versions of this manuscript. Parts of this work have been supported by the FP7 Collaborative Project MICROME (grant agreement # 222886-2), funded by the European Commission. C.A.O. thanks the Department of Biological Sciences at the University of Cyprus for their kind hospitality during the spring semester of 2012.

REFERENCES

AAAS, 2005. *The State of Major Ecosystems*. Available at: http://atlas.aaas.org/index.php?part=1&sec=status.

answers.com, 2005. *Large Numbers*. Available at: http://www.answers.com/topic/large-numbers.

Bryson B, 2003. *A short history of nearly everything.* Broadway Books, New York.

Burgess SC, 2004. Proteomics in the chicken: tools for understanding immune responses to avian diseases. *Poultry Science,* 83: 552-573.

Chothia C, 1992. Proteins. One thousand families for the molecular biologist. *Nature,* 357: 543-544.

Cirkovic MM, 2004. The temporal aspect of the Drake equation and SETI. *Astrobiology,* 4: 225-231.

Cornish-Bowden A, Cardenas ML, 2005. Systems biology may work when we learn to understand the parts in terms of the whole. *Biochemical Society Transactions,* 33: 516-519.

Cowley AW, 2004. The elusive field of systems biology. *Physiological Genomics,* 16: 285-286.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW, 2006. The evolution of mammalian gene families. *PLoS One,* 1: e85.

Drake F, 2004. The E.T. equation, recalculated. *Wired,* 12: 225.

Goodman L, 2003. *Making a Genesweep: It's Official!* Available at: http://www.bio-itworld.com/archive/071503/genesweep.

Grigoriev A, 2003. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research,* 31: 4157-4161.

Hook PE, 2004. *A Drake equation for linguistic diversity*. Available at: http://www-personal.umich.edu/~pehook/drake.html.

Ideker T, Galitski T, Hood L, 2001. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2: 343-372.

Karev GP, Wolf YI, Berezovskaya FS, Koonin EV, 2004. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology*, 4: 32.

Kunin V, Cases I, Enright AJ, de Lorenzo, Ouzounis CA, 2003. Myriads of protein families, and still counting. *Genome Biology,* 4: 401.

Kunin V, Teichmann SA, Huynen MA, Ouzounis CA, 2005. The properties of protein family space depend on experimental design. *Bioinformatics,* 21: 2618-2622.

Lotka AJ, 1925. *Elements of physical biology*. Williams & Wilkins Co, Baltimore.

May M, 2005. *Systems biology: interdisciplinary integration and beyond*. Available at: http://www.highbeam.com/doc/1G1-131128560.html.

Naylor S, Cavanagh J, 2004. Status of systems biology does it have a future? *Drug Discovery Today - Biosilica,* 2: 171-174.

nso.edu, 2001. *Mass, Size, and Density of the Universe*. Available at: http://people.cs.umass.edu/~immerman/stanford/universe.html.

Ouzounis CA, Coulson RMR, Enright AJ, Kunin V, Pereira-Leal JB, 2003. Classification schemes for protein structure and function. *Nature Reviews Genetics*, 4: 508-519.

Peregrin-Alvarez JM, Parkinson J, 2007. The global landscape of sequence diversity. *Genome Biology,* 8: R238.

Pereira-Leal JB, Enright AJ, Ouzounis CA, 2004. Detection of functional modules from protein interaction networks. *Proteins - Structure Function and Genetics,* 54: 49-57.

Rabinowicz PD, Vollbrecht E, May B, 2000. How many genes does it take to make a human being? *Genome Biology,* 1: reports4013.1-4013.3.

Sammut SJ, Finn RD, Bateman A, 2008. Pfam 10 years on: 10,000 families and still growing. *Briefings in Bioinformatics,* 9: 210-219.

Spirin V, Mirny LA, 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100: 12123-12128.

Tautz D, Domazet-Loso T, 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics,* 12: 692-702.

Todar K, 2008. *The bacterial flora of humans*. Available at: http://textbookofbacteriology.net/normalflora.html.

Veizer J, 1988. The earth and its life: systems perspective. *Origin of Life and Evolution of Biospheres*, 18: 13-39.

wikipedia.org, 2012. *Drake equation - current estimates*. Available at: http://en.wikipedia.org/ wiki/Drake_equation#Current_estimates.

Yang X, Xie L, Li Y, Wei CC, 2009. More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. *PLoS One,* 4: e6074.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li WZ, *et al.*, 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology,* 5: 432-466.