BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2918–2924
doi:10.1093/bioinformatics/btt505

Data and text mining

Advance Access publication August 31, 2013

Quantifying the complexity of medical research

Raul Rodriguez-Esteban* and William T. Loging

Computational Biology, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT 06877, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: A crucial phenomenon of our times is the diminishing marginal returns of investments in pharmaceutical research and development. A potential reason is that research into diseases is becoming increasingly complex, and thus more burdensome, for humans to handle. We sought to investigate whether we could measure research complexity by analyzing the published literature.

Results: Through the text mining of the publication record of multiple diseases, we have found that the complexity and novelty of disease research has been increasing over the years. Surprisingly, we have also found that research on diseases with higher publication rate does not possess greater complexity or novelty than that on less-studied diseases. We have also shown that the research produced about a disease can be seen as a differentiated area of knowledge within the wider biomedical research. For our analysis, we have conceptualized disease research as a parallel multi-agent search in which each scientific agent (a scientist) follows a search path based on a model of a disease. We have looked at trends in facts published for diseases, measured their diversity and turnover using the entropy measure and found similar patterns across disease areas.

Contact: raul.rodriguez-esteban@roche.com

Received on April 12, 2013; revised on August 26, 2013; accepted on August 27, 2013

1 INTRODUCTION

Diseases are complex problems investigated by myriads of scientists working in parallel. Most of these scientists work on narrow niches to produce pieces of knowledge that contribute to our overall understanding of diseases. This understanding then enables the test of therapeutic interventions that might be better alternatives to existing treatments. To investigate this process, we have used the analogy of a multi-agent system throughout this study. We compare disease research with a search performed by a multi-agent system in which the agents (scientists) explore a space (a disease) to enable the selection of the best solutions (interventions) to a problem. This analogy has limitations, but it aids in delving into the intrinsic properties of the problem (the metaproblem). There are similar motivations in intra-human modeling that can be seen in projects such as the Virtual Physiological Human (Hunter *et al.*, 2010) or ApiNATOMY (de Bono *et al.*, 2012).

The search agents in our multi-agent system use models of reality to guide their path. These models represent the models that scientists use to guide their research. An important part of disease research involves building models that recapitulate some of the main properties of diseases. The best known of these models are *in vitro* and *in vivo* proxies that are lower-cost lower-risk alternatives to clinical studies (e.g. the unilateral ureteral obstruction mouse is an animal model for human kidney disease). There are also disease models of a mathematical nature that represent diseases at the systems or epidemiological level. Another type of model is based on the knowledge available for a disease (Thagard, 2012), such as the models used in knowledge-based clinical decision support systems. Biomedical researchers use knowledge models of diseases (Coden *et al.*, 2009) based on knowledge acquired from textbooks, scientific literature and experimental and clinical experience. They are made up of multiple elements, which can be clinical, cellular, molecular and public health related—to name a few. These models are the focus of this study.

For each disease, researchers pay more attention to different sets of elements. For example, clinicians may use a patient's symptoms, demographics, clinical history and laboratory reports to evaluate disease progression. Biologists, on the other hand, may use genes, pathways, organelles and other elements to generate hypotheses about the mechanism of action of a drug or the activity of signaling pathways. In a disease such as psoriasis, a clinician may focus on the relationship between psoriasis area and severity index (PASI) scores and different immunosuppressive drugs, whereas a biologist may investigate the interplay of cutaneous and immune cells. In diabetic nephropathy, on the other hand, the interest may fall on the biology of podocytes and myofibroblast proliferation or on a patient's glomerular filtration rate and arterial blood pressure. Thus, different elements are combined by each scientist to construct a model of reality that enables reasoning about the disease. We incorporate such scientific models in our multi-agent system analogy by postulating that each scientific agent has a model of a disease built into its internal state. This model helps the agent navigate its path and is based on inputs produced by other agents, such as scientific publications, compounds or tools. These models are crucial because the agents cannot perform their search through brute-force exploration. They need to choose a path based on the likelihood of success. Thus, disease models facilitate decision making.

We were interested in learning how the structure of these disease models can affect the task of the search agents. We started with the hypothesis that some disease models might be more complex than others for several reasons. One of the reasons could come from the number of articles published about a disease per year, as scientific publishing has been increasing exponentially across many areas of medical research (Larsen and von Ins, 2010). Another reason could come from anatomy. For example, psoriasis is a fairly localized disease of the skin;

*To whom correspondence should be addressed.

however, a disease such as diabetes involves multiple organs and tissues. Yet another reason could come from the multiplicity of signaling pathways involved in a disease. Thus, a greater multiplicity of elements in a disease could mean a greater search space to investigate.

A theoretical disease model can further illustrate this latter point. This theoretical model involves a non-existent animal species called *bigenes*, which has two genes in its genome: *A* and *B*. *Bigenes* suffers a species-specific disease called X, which is known to be driven by a mutation in *A*. Up to the year *t*, the literature on X comprised five articles focused on gene *A*. The publication record about X at time *t* could be represented as $X_t = \{AAAAA\}$. Thus, X_t is a simplified representation of what was known about X up to time *t*. The genes belonging to X_t may not be disease genes in a traditional sense, but they represent the scientific conversation taking place about X, even in cases when negative results are reported, which may be contradicted later on. Defining 'disease genes' is in itself a complicated task as the relationship between a gene and a disease can range from strong and firmly established to tenuous, depending on the weight of the existing evidence.

A breakthrough in the research of X occurred in the year $t+1$, when a mutation in *B* was newly associated with X. New articles appeared investigating this association; however, other articles still focused on *A*, leading to $X_{t+1} = \{AAAAABABBA\}$, where the subsequence $\{BABBA\}$ corresponds to publications for the year $t+1$. Thus, the discovery of a mutation in *B* made the scientific model of the disease X more complex, and this is reflected in the publication record. This change in complexity can be quantified using the entropy measure from information theory, which is a common measure of statistical complexity (Feldman and Crutchfield, 1998) used in biomedical applications such as monitoring heart rates (Perkiömäki *et al.*, 2005) and brain electrical activity (Stam, 2005) or in evolutionary biology (Adami, 2002). The entropy for X_t is $H(X_t) = 0$, but the introduction of *B* in year $t+1$ leads to $H(X_{t+1}) = 0.88$ (see Section 2). $H(X_{t+1}) > H(X_t)$ reflects the increase in complexity over time.

A number of works have been devoted to analyzing the growth and evolution of publications on gene and protein interactions outside the context of disease by analyzing the accumulated published record (Arbesman and Christakis, 2011; Cokol *et al.*, 2005; Cokol and Rodriguez-Esteban, 2006; He and Zhang, 2009; Hoffmann and Valencia, 2003; Pfeiffer and Hoffmann, 2007). However, there is a divergence between the published record and the focus of active research, as older works fade in influence (on average). We believe that practical disease models do not involve every piece of knowledge published about a disease. Defining boundaries, nonetheless, is not straightforward. In the case of the disease X, publications were initially only about gene *A*, but that changed after an important discovery. Realizing this requires specific knowledge of the disease and a scientific consensus that may not be found. To generalize, we have used a time window. A model containing the state of the art in year $t+1$ for a 1-year time window would be $X'_{t+1} = \{BABBA\}$. We call this a *fading memory* model, and it reflects that the older a finding is, the less relevant it becomes—just like a fading memory. This decay can be observed in the evolution of citation counts of published literature over time.

One shortcoming of the fading memory approach is that some diseases have much higher publication rate than others. However, memories may fade faster for diseases with higher publication rate owing to the finite memory of the scientists that study them. A *finite memory* approach would consider the latest *n* publications rather than a time window. In the case of X, this would mean a disease model $X'_{t+1} = \{BA\}$ for a memory size of $n = 2$. We have centered our analysis on the finite memory approach because publication rates change widely over time—and between diseases—and thus comparing fixed time windows would require some type of normalization, which could introduce distortions. Nonetheless, the fading memory approach is also discussed, as will be shown later in the text.

Another important aspect of disease models is their evolution in time. A rapid pace of change may signify, for example, that the principles of a disease are not properly defined. A disease model whose elements have a high turnover rate requires a greater effort at understanding and leaves less time to devote to its details. In the case of the disease X, the appearance of *B* in the publication record was unexpected, which in Bayesian probability terms can be expressed as $p(B \in X'_{t+1} \mid X'_t) = 0$. On the other hand, articles about *A* were fully anticipated, $p(A \in X'_{t+1} \mid X'_t) = 1$. Thus, conditional probabilities can be used to gauge the rate of change in a disease model and the novelty of a finding.

The *bigenes* example is about an organism with only two genes (like some simple viruses), but a similar case could have been made about an organism with only two cell types because research emphasis on cell types also changes over time. For example, much immunological research has been focused on the T-cell subsets Th_1 and Th_2, but, recently, there has been increased emphasis on Th_{17} and T_{reg} cells, whereas new cell types, such as Th_9 and Th_{22}, are regularly proposed. In this analysis, we have considered disease models not only involving genes (as in the *bigenes* example) but also cell types, drugs and chemicals, and we have looked at how they influence each other.

2 METHODS

Disease literature data came from Medline 2012 release abstracts annotated with at least one disease MeSH term (branch C of the MeSH Tree Structures, which covers human and animal diseases) and published between 1970 and 2010. Disease groups were based on the main subbranches of the disease MeSH tree. Gene, cell type, drug and chemical annotations came from GeneView [(Thomas *et al.*, 2012); download July 2012]. GeneView uses the GNAT (Hakenberg *et al.*, 2011) algorithm for gene recognition, which showed 82% precision and 82% recall for abstracts in the BioCreative challenge, and which has been extended to 20 model species. Chemical entity annotations in GeneView come from ChemSpot (Rocktäschel *et al.*, 2012), which achieves 68% precision and 69.5% recall. Cell types and drugs are based on the AliBaba (Plake *et al.*, 2006) dictionaries of regular expressions and spelling variations.

For each disease in the MeSH Tree Structures, an ordered sequence was created $X_k = \{x_{k,1}, x_{k,2}, x_{k,3}, \ldots\}$, where each $x_{k,i}$ is the name of a gene mentioned in an abstract annotated with the MeSH term corresponding to disease *k*. Each sequence X_k was ordered by publication time. That is, for every $x_{k,i}$, there is an associated time stamp $t_{k,i}$ for which $t_{k,i} \leq t_{k,j}$ if $i \leq j$. The time stamp was based on the publication date of the abstract. For publication dates that only specify month and/or year, the earliest possible date was assigned (e.g. the first day of the month). For publications appearing on the same date, the PubMed ID value was used

to decide order. The X_k sequences are, therefore, time series that can be analyzed and compared.

To apply the *finite memory* approach, we considered every possible ordered sequence $X_k^j \in X_k$ of length n, $X_k^j = \{x_{k,j}, x_{k,j+1}, \ldots, x_{k,j+n-1}\}$. Thus, each X_k^j is a disease model of disease k under the finite memory approach for a memory of size n. The initial value of n chosen was 100. Overall conclusions did not change for other values of n tested. Entropy was computed for each X_k^j, following Shannon's theorem:

$$H(X_k^j) = -\sum_l p_k^j(A_l)\log(p_k^j(A_l)), \quad (1)$$

where $p_k^j(A_l)$ is the frequency of gene A_l within the sequence X_k^j or

$$p_k^j(A_l) = \frac{\sum_{i\in(j,j+n-1)} I_{x_{k,i}=A_l}}{n}, \quad (2)$$

where $I_{x_{k,i}=A_l}$ is a binary function that returns one when $x_{k,i}$ is equal to A_l, and zero otherwise. Compare this with the log-entropy models for information retrieval in (Roy *et al.*, 2011).

Each $H(X_k^j)$ was associated to a time stamp $\bar{t}_k^j$, which is the average of the time stamps associated to the $x_{k,j}$ components of X_k^j. The publication rate for each X_k^j was computed as follows:

$$r_k^j = \frac{n}{t_{k,j+n-1} - t_{k,j}}. \quad (3)$$

To assess the evolution of $H(X_k^j)$ over time and publication rate, $H(X_k^j) \sim f(\bar{t}_k^j, r_k^j)$, two methods were used, which produced similar overall picture. The first method involved computing for each disease k, the multiple linear regression between the entropy and the publication time and publication rate $\{H(X_k^j), \bar{t}_k^j, r_k^j\}$ for all j values. The second method involved creating a random sample of values picked among all the $\{H(X_k^j), \bar{t}_k^j, r_k^j\}$ triplets. Randomization followed these steps: a disease k was randomly picked and then a random value of j for that disease was selected. This sampling method was implemented to make sure that diseases with longer sequences did not dominate the results. The random sample had size 100 000.

To measure the novelty of new findings, the conditional probability of an element $x_{k,j+n}$ for each sequence X_k^j was computed as follows:

$$p_k^j(x_{k,j+n}|X_k^j) = \frac{\sum_{i\in(j,j+n-1)} I_{x_{k,i}=x_{k,j+n}}}{n} \quad (4)$$

and then compared with the time and rate of publication, $p_k^j(x_{k,j+n}) \sim f(\bar{t}_k^j, r_k^j)$. To smooth the results, the average of a forward window of $w_0 = 10$ elements was used:

$$\frac{\sum_{w\in(0,1,\ldots,w_0-1)} p_k^j(x_{k,j+n+w}|X_k^j)}{w_0} \quad (5)$$

Control sequences similar to X_k were created by shuffling the disease annotations across the corpus (Fisher–Yates random permutation). Using this method, many of the properties of the dataset, such as the overall frequency of genes and diseases, are preserved. This represents a higher bar than alternatives such as generating random disease and/or gene annotations from a uniform distribution.

The same process was followed with the annotations of chemicals and drugs. For cell types, however, the data were insufficient to construct X_k^j memory sequences of length 100, and thus length 20 was used instead. Thus, some of the results presented for cell types are not completely comparable with the rest. Correlations between the $H(X_k^j)$ of genes, chemicals, drugs and cell types for each disease k were performed using entropy sequences with monthly values based on linear interpolation of the values of $H(X_k^j)$.

3 RESULTS

The complexity of disease models made of genes has been increasing over time, as can be seen in Figure 1. This is reflected

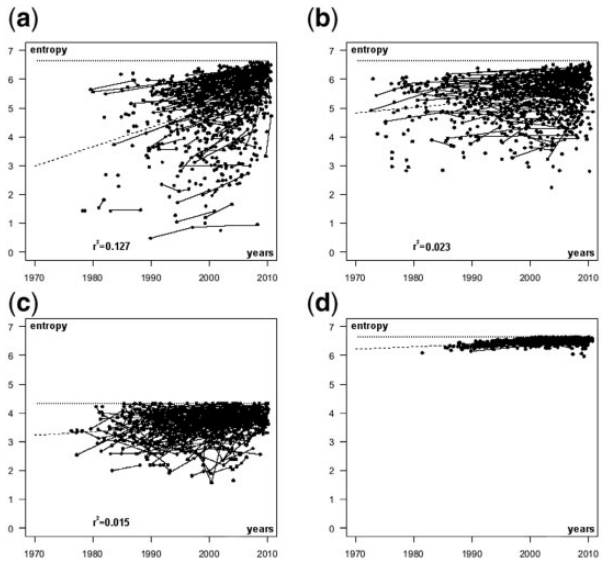

Fig. 1. Entropy of disease models over time. Points represent randomly picked samples; edges connect points from the same disease. The dotted line is the maximum possible entropy. The dashed line is a linear regression. (**a**) Entropy of gene disease models. (**b**) Entropy of chemical disease models. (**c**) Entropy of cell disease models. (**d**) Entropy of control gene disease models. Maximum entropy was calculated as $\log_2(n)$, where n is the size of the analysis window. The value of n was 100 for (a), (b) and (d), and 20 for (c)

in the r^2 and the multiple linear regression coefficients in Table 1. To test the robustness of the result, we changed the value of the time window (originally $n = 100$) to $n \in \{25, 50, 75, 100, 125, 150, 175, 200\}$ and found that the relation persisted and even became stronger with increasing n. The r^2 fit increased with the size n of the window ($r^2 \sim 3.4 \times 10^{-2} \ln(n) - 2.8 \times 10^{-2}$), as did the coefficients for time ($r^2 \sim 2.3 \times 10^{-2} \ln(n) - 4.2 \times 10^{-2}$) and publication rate ($r^2 \sim 1.1 \times 10^{-5} \ln(n) - 3.4 \times 10^{-5}$). Choosing the value of n is a trade-off between resolution and coverage, as diseases with less than n published facts are not included in the analysis. We also looked within disease groups and found similar patterns (see Fig. 2 and Table 2), with, for example, cardiovascular diseases and nervous system diseases increasing in complexity at a faster pace over the period involved, and neoplasms and substance-related disorders growing slower.

An illustrative example is the disease retinitis pigmentosa (RP), a retinal degenerative disease. The entropy of its disease model over time fits a linear regression with $r^2 = 0.89$ and a positive slope of 0.13, indicating that its complexity has been increasing over time (see Fig. 2d). The lower early entropy is illustrated by one of the main genes studied in RP: rhodopsin. Owing to the early discovery of a rhodopsin mutation associated to RP (Dryja *et al.*, 1990, 1991), rhodopsin was important in publications about RP, leading to initial low entropy in the RP disease model. However, rhodopsin has faded relatively in importance as many other mutations and genes have been associated to RP in recent times. The current multiplicity of genes studied has led to higher entropy in the disease model of RP.

For disease models of chemicals and drugs, we can see a milder increase in complexity (see Section 2). Thus, the increase in

Table 1. Multiple linear regression coefficients and r^2 for the entropy of genes, chemicals, drugs and cell types with respect to time (years) and publication rate (publications with findings per year)

Category	Time	Rate	r^2	$P<$
Genes	6.44×10^{-2}	1.12×10^{-5}	0.127	1×10^{-15}
Chemicals	1.27×10^{-2}	4.86×10^{-6}	0.023	7.99×10^{-13}
Drugs	1.18×10^{-2}	7.85×10^{-6}	0.013	4.05×10^{-7}
Cells	5.20×10^{0}	-1.72×10^{-3}	0.015	1.29×10^{-8}

Note: Category refers to the type of disease model. Time and rate are the coefficients for the multiple linear regression for the disease model in category. Adjusted r^2 is shown for the overall regression. *P*-values for time coefficients are $<10^{-15}$, and *P*-values for publication rate coefficients are at least $<10^{-9}$. As explained in Section 2, values for the cell type category are not directly comparable with the rest.

Table 2. Multiple linear regression coefficients and r^2 for the entropy of genes with respect to time (years) and publication rate (publications with findings per year) for different disease groups

Category	Time	Rate	r^2
Cardiovascular diseases	8.32×10^{-2}	3.93×10^{-7}	0.336
Nervous system diseases	8.16×10^{-2}	6.43×10^{-6}	0.208
Endocrine system diseases	7.19×10^{-2}	1.25×10^{-5}	0.252
Female urogenital diseases and pregnancy complications	6.93×10^{-2}	2.24×10^{-5}	0.267
Congenital, hereditary and neonatal diseases and abnormalities	6.88×10^{-2}	1.64×10^{-4}	0.138
Animal diseases	6.74×10^{-2}	3.85×10^{-6}	0.168
Bacterial infections and mycoses	6.51×10^{-2}	1.74×10^{-4}	0.119
Digestive system diseases	5.82×10^{-2}	6.56×10^{-6}	0.242
Parasitic diseases	5.36×10^{-2}	2.76×10^{-5}	0.397
Musculoskeletal diseases	5.07×10^{-2}	2.00×10^{-5}	0.128
Virus diseases	4.99×10^{-2}	2.00×10^{-7}	0.268
Otorhinolaryngologic diseases	4.38×10^{-2}	1.19×10^{-3}	0.235
Respiratory tract diseases	4.27×10^{-2}	6.67×10^{-6}	0.154
Male urogenital diseases	4.16×10^{-2}	1.18×10^{-5}	0.136
Nutritional and metabolic diseases	4.15×10^{-2}	1.04×10^{-5}	0.085
Eye diseases	4.10×10^{-2}	1.29×10^{-4}	0.147
Hemic and lymphatic diseases	4.00×10^{-2}	1.94×10^{-5}	0.095
Pathological conditions, signs and symptoms	3.72×10^{-2}	5.35×10^{-6}	0.090
Immune system diseases	3.66×10^{-2}	3.88×10^{-6}	0.181
Neoplasms	3.19×10^{-2}	3.93×10^{-6}	0.147
Skin and connective tissue diseases	3.07×10^{-2}	6.45×10^{-7}	0.086
Stomatognathic diseases	3.07×10^{-2}	2.99×10^{-4}	0.041
Substance-related disorders	2.68×10^{-2}	2.69×10^{-6}	0.123
Occupational diseases	2.15×10^{-2}	2.82×10^{-3}	0.112

Note: Category refers to the group of diseases covered by the disease model. Time and rate are the coefficients for the multiple linear regression for the disease model in category. Adjusted r^2 is shown for the overall regression. *P*-values for time coefficients are $<10^{-15}$ and for rate coefficients are at least $<10^{-6}$, except for virus diseases, which are not significant.

complexity is not homogeneous across all categories of elements. This is seen in models of RP as well. The linear regression for the disease model of chemicals has r^2 of 0.54 and slope of 0.06, and for drugs r^2 of 0.80 and slope of 0.05. The low entropy in

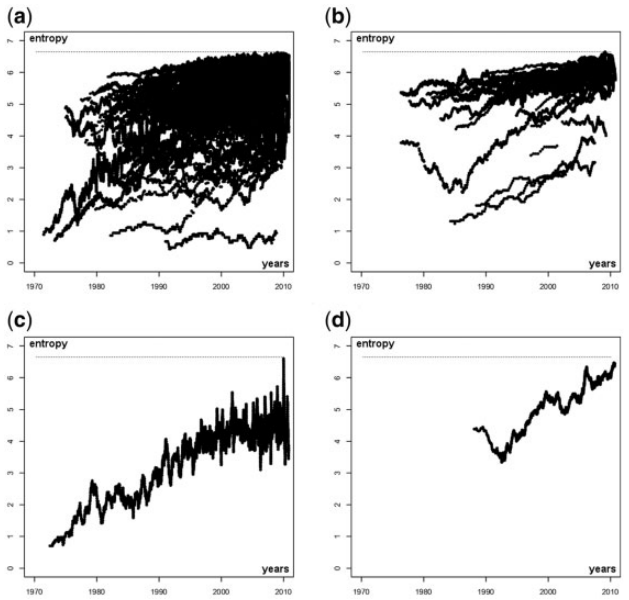

Fig. 2. Entropy of gene disease models over time for selected disease groups and diseases. The dotted line is the maximum possible entropy. The dashed line is a linear regression. (**a**) Entropy for the disease group cardiovascular diseases. (**b**) Entropy for the disease group substance-related disorders. (**c**) Entropy for the disease multiple sclerosis. (**d**) Entropy for the disease RP. Maximum entropy was calculated as $\log_2(n)$, where n is the size of the analysis window, $n = 100$

chemical models in the earlier studies of RP is illustrated by the relatively higher prominence of fluorescein, whose use tended to be highlighted in the abstract [e.g. (Best *et al.*, 1971)], a practice which has faded since. Fluorescein is counted both as a chemical and as a drug (FDA approved in this case). Thus, the drug and chemical categories overlap.

The control disease models created by random permutation of disease annotations (see Section 2) do not show any of these properties (see Fig. 1d). Moreover, control disease models do not show the heterogeneity of entropy values seen in actual disease models, suggesting that entropy differences are significant between disease models. For example, the linear regression that fits the control gene disease model for RP has r^2 and slope close to 0 and does not show dependence with time. Moreover, the entropy values are high and close to what a discrete uniform distribution with sufficient cardinality would produce. Actual disease models have lower entropy than control disease models because they are concentrated around certain elements such as popular genes or drugs and form differentiated areas of knowledge, as the RP example shows.

An implication is that 100 randomly selected articles about a disease should, on average, cover less unique facts (e.g. genes) than 100 articles picked at random from the overall literature. This means that the focus of a disease's research is narrower and there is more repetition, which gives it more coherence and makes it more understandable to humans. When the focus of a disease becomes too broad it could be an indication that the disease definition has become too inclusive.

As seen in the disease models of RP, the slope of the linear regression between entropy and time tends to be positive,

indicating that entropy increases over time. For gene disease models, the median slope is 0.04184, for chemicals 0.01269 and for drugs 0.0125 (t-tests showed these values to be significantly >0). The interquartile ranges of the slope values are largely positive: [0.01668, 0.0746] for genes, [−0.00152, 0.03049] for chemicals and [0.0125, 0.03273] for drugs. Looking at all diseases, we found these values to be unrelated to the average publication rate, showing that disease models with high average publication rates are no more complex than those of diseases with low average publication rates. This is similarly reflected in Table 1, where we show that a small fraction of the increase in complexity is driven by higher publication rates. For example, in gene disease models, the linear regression coefficients indicate that a 5000-fold publication rate increase would increase the complexity as little as 1 year of time would.

By looking at the RP example alone, we could not have known whether the increase in entropy of its disease models was somehow related to the increase in the rate of publications about RP that happened over the same time period. Our analysis, however, allows us to say that these two variables are generally unrelated. Instead, diseases with high publication rates would seem to have a 'fast forward' behavior in which findings are brought up faster but without an increase in associated complexity. Thus, the *finite memory* approach chosen here (as described in Section 2) seems more appropriate to study this phenomenon than the *fading memory* approach.

As can be seen in Figure 3 and Table 3, the novelty in disease models made of genes has been increasing over time as well. This can be seen in the negative relation between time and the probability of a publication mentioning a gene that is already in a disease model. For chemicals, this relationship is barely significant, and for drugs, it is not significant. The linear regression for disease models of genes has a median slope of −0.00942 and an interquartile range of [−0.03115, −0.00101] (here, negative slope means increased novelty). Increased novelty could be one of the main drivers of increased entropy, as they are related. In the gene disease model of RP, novelty increases noticeably with time, and fits a linear regression of slope −0.07 and r^2 of 0.63. From the complexity and novelty of the disease models of RP, we postulate that RP has become harder to study over time. RP is not a unique case, as can be seen in Figure 2 for two disease groups and multiple sclerosis.

As shown in Table 3, we found that publication rates do not correlate with novelty. Thus, a slower pace of publication does not entail a lower appetite for novelty in less-explored diseases. Control disease models show a more stable and much higher rate of novelty, as would be expected from a randomly shuffled set.

To further study the drivers of complexity in disease models we measured the correlations between the entropies of disease models for genes, chemicals and drugs (see Section 2). These correlations were found to be positive (see Fig. 4 and Table 4), meaning that increases in one category were correlated with increases in another category, even controlling for the common increase of entropy over time already described. In the case of RP, for example, the entropies for the disease models of genes and chemicals were correlated, with a slope of 0.50 and r^2 of 0.58. This shows that changes in complexity of disease models are not independent for each category and that it would be reasonable to consider disease models in a multidimensional fashion, with different categories of elements interacting with each other rather than studying them separately. It should also be noted that variations in complexity for each category may have different practical impact. For example, an increase in cell type complexity could be more relevant than an increase in gene complexity because a gene can be associated to different functions in different cell types.

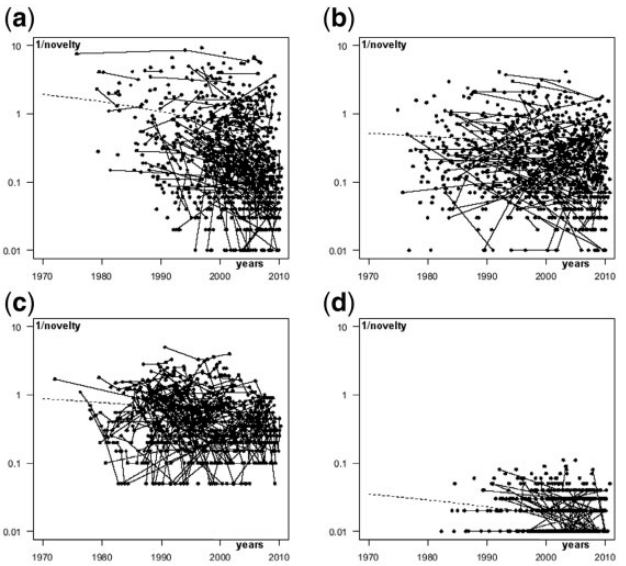

Fig. 3. Novelty of disease models over time. Smaller values on the y-axis represent higher novelty. Points are randomly picked samples, edges connect points from the same disease. The dashed line is a linear regression between time and novelty. (**a**) Novelty of gene disease models. (**b**) Novelty of chemical disease models. (**c**) Novelty of cell disease models. (**d**) Novelty of control gene disease models

Table 3. Multiple linear regression coefficients and r^2 for the novelty of genes, chemicals, drugs and cells with respect to time (years) and publication rate (publications with findings per year)

Category	Time	Rate	r^2	$P<$
Genes	-4.20×10^{-2}	-2.89×10^{-5}	0.057	1.08×10^{-13}
Chemicals	-6.27×10^{-4}	-4.63×10^{-6}	0.006	1.57×10^{-2}
Drugs	-5.22×10^{-3}	-9.86×10^{-5}	−0.001	7.33×10^{-1}
Cells	-1.14×10^{-2}	8.84×10^{-4}	0.024	2.07×10^{-6}

Note: P-values for time coefficients are not significant for drugs, and P-values for publication rate coefficients are only significant for cell types. Adjusted r^2 is shown for the overall regression. As explained in Section 2, values for the cell category are not directly comparable with the rest.

4 CONCLUSIONS

Our study shows quantitatively that the complexity of studying diseases has been increasing over time. A resulting implication is that selecting molecular targets in target-based drug discovery becomes a more arduous task as the search space grows,

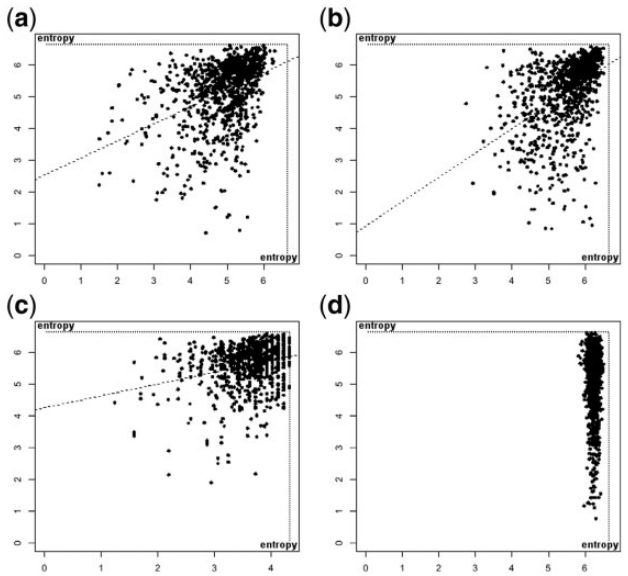

Fig. 4. Correlation between entropies. Points represent randomly picked samples. The dashed line is a linear regression. In all four graphs, the *y*-axis represents gene disease model entropy. (**a**) Correlation between gene disease model entropy and drug disease model entropy. (**b**) Correlation between gene disease model entropy and chemical disease model entropy. (**c**) Correlation between gene disease model entropy and cell disease model entropy. (**d**) Correlation between gene disease model entropy and control gene disease model entropy

Table 4. Correlation between entropy values for different categories of disease models

Category 1	Category 2	Correlation	Time	r^2	$P<$
Genes	Drugs	6.57×10^{-1}	3.98×10^{-2}	0.260	1×10^{-15}
Genes	Chemicals	5.77×10^{-1}	3.68×10^{-2}	0.218	1×10^{-15}
Chemicals	Drugs	4.99×10^{-1}	2.64×10^{-3}	0.403	1×10^{-15}

Note: All coefficients have associated *P*-values $<10^{-15}$, except for the coefficient for time when comparing chemicals and drugs, which is not significant. The correlation values are between category 1 and category 2.

especially given the limitations of existing therapeutic tools (Rodriguez-Esteban *et al*., 2009). Pharmaceutical companies need to scrutinize and triage more genes to find the best candidate for development. The effort increases if the set of genes being studied changes quickly over time—and the same applies to cells, tissues and compounds. Further studies are necessary to ascertain the link between this complexity and increased R&D costs.

The larger the diversity also means the harder the problem becomes for the limited human cognitive abilities. This can be seen in a more simplified multi-agent search example. In this setting, scientific production about a disease comes from agents that only study and publish about a single gene, without consideration of other genes. These agents publish—all of them—at exactly the same rate. Some diseases have higher publication rates, but this is as a result of a larger number of agents working on them owing to greater perceived significance. In this scenario, it would be possible for those diseases with higher publication rate to have more complex models because each agent can specialize in a different gene and create a separate research niche. Thus, having more agents at work would allow for more bandwidth on the production side, particularly 'brain bandwidth'.

However, we have shown that diseases with higher publication rate do not have more complex disease models. One explanation could be that scientists are also consumers of the research produced by other scientists, and therefore they naturally tend to follow the path that others have opened (Cokol *et al*., 2005)—in some cases because it is the only path being funded. Another explanation is that individual scientists can only process a certain level of complexity within the limits of their cognitive abilities and the technologies available to them. Hence, there would be a limitation on the consumption side of science, and scientific production would tend to adapt its output to this consumption 'bandwidth'. These explanations would leave the open question of why disease complexity and novelty have, nonetheless, been increasing over time, as we have shown. A potential reason is that technological progress has brought forward high-throughput experimental and *in silico* technologies that enhance the ability of scientists to produce, process and understand large datasets.

Although genetic screens and other unbiased high-throughput techniques are less affected by an increase in the multiplicity of prior knowledge, these experiments usually produce extensive new results that need to be validated with low-throughput methods by researchers grappling with them (Mons and Velterop, 2009). We should also note that many high-throughput techniques are not completely unbiased and depend on prior knowledge. For example, the therapeutic usefulness of a cell-based screen will depend on the choice of cell type.

Repeated failure in clinical trials has led to increased efforts at patient stratification, using such techniques as genotyping or gene expression analysis. Under the framework presented here, stratification could be understood as a way of drawing more convenient boundaries to a problem to reduce its complexity. For example, the diversity found in cancer has led to the proposition that it should be considered a constellation of different diseases classified by organs, cell types, molecular profiles, driver somatic mutations, active pathways and so forth, which could be addressed with computational and text mining approaches (Clancy *et al*., 2011). Slicing a problem can be an effective way to simplify it.

Besides the factors considered here, there are other reasons why some diseases should be more difficult to investigate, such as experimental limitations (e.g. lack of satisfactory animal models or cell cultures such as in neurological disorders, imprecise data from high-throughput techniques) or clinical constrains (e.g. frequent comorbidities, inadequate biomarkers, uncertain diagnostic tools). Genetic or phenotypic heterogeneity or multiple associated genetic loci, may hinder the understanding of a disease's genetic underpinnings. Other difficulties could be driven by disease-protective mechanisms, such as compensatory feedbacks, fast mutation rates or redundant pathways that make diseases robust against therapeutic intervention. Moreover, literature biases such as unpublished negative results or published

conclusions that are unsupported by the available data, as well as heterogeneity in the quality of the data, could interplay with the complexity we are measuring or add another layer to the problem. Beyond these scientific factors, non-scientific ones such as regulatory and legislative problems or patenting issues can loom large for certain diseases.

Ultimately, complexity should be considered primarily as causing a tangible impact to the input factors necessary for scientific research (time, money, patients, etc.). In this light, the need for better understanding is highly relevant, considering the stagnating pace of pharmaceutical research despite the abundance of incentives to improve existing therapies.

ACKNOWLEDGEMENTS

We would like to thank Rohitha P. SriRamaratnam for comments on the article.

Conflict of Interest: none declared.

REFERENCES

Adami,C. (2002) What is complexity? *Bioessays*, **24**, 1085–1094.

Arbesman,S. and Christakis,N.A. (2011) Eurekometrics: analyzing the nature of discovery. *PLoS Comput. Biol.*, **7**, e1002072.

Best,M. *et al.* (1971) Fluorescein angiography during induced ocular hypertension in retinitis pigmentosa. *Am. J. Ophthalmol.*, **71**, 1226–1230.

de Bono,B. *et al.* (2012) ApiNATOMY: a novel toolkit for visualizing multiscale anatomy schematics with phenotype-related information. *Hum. Mutat.*, **33**, 837–848.

Clancy,T. *et al.* (2011) Immunological network signatures of cancer progression and survival. *BMC Med. Genomics*, **4**, 28.

Coden,A. *et al.* (2009) Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J. Biomed. Inform.*, **42**, 937–949.

Cokol,M. *et al.* (2005) Emergent behavior of growing knowledge about molecular interactions. *Nat. Biotechnol.*, **23**, 1243–1247.

Cokol,M. and Rodriguez-Esteban,R. (2006) Visualizing evolution and impact of biomedical fields. *J. Biomed. Inform.*, **41**, 1050–1052.

Dryja,T.P. *et al.* (1990) Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. *N. Engl. J. Med.*, **323**, 1302–137.

Dryja,T.P. *et al.* (1991) Mutation spectrum of the rhodopsin gene among patients with autosomal dominant retinitis pigmentosa. *Proc. Natl Acad. Sci. USA*, **88**, 9370–9374.

Feldman,D.P. and Crutchfield,J.P. (1998) Measures of statistical complexity: why? *Phys. Lett. A*, **238**, 244–252.

Hakenberg,J. *et al.* (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

He,X. and Zhang,J. (2009) On the growth of scientific knowledge: yeast biology as a case study. *PLoS Comput. Biol.*, **5**, e1000320.

Hoffmann,R. and Valencia,A. (2003) Life cycles of successful genes. *Trends Genet.*, **19**, 79–81.

Hunter,P. *et al.* (2010) A vision and strategy for the virtual physiological human in 2010 and beyond. *Philos. Trans. A Math. Phys. Eng. Sci.*, **368**, 2595–614.

Larsen,P.O. and von Ins,M. (2010) The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, **84**, 575–603.

Mons,B. and Velterop,J. (2009) Nano-publication in the e-science era. In: *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)* CEUR Workshop Proceedings, Washington, DC, USA.

Perkiömäki,J.S. *et al.* (2005) Fractal and complexity measures of heart rate variability. *Clin. Exp. Hypertens.*, **27**, 149–158.

Pfeiffer,T. and Hoffmann,R. (2007) Temporal patterns of genes in scientific publications. *Proc. Natl Acad. Sci. USA*, **104**, 12052–12056.

Plake,C. *et al.* (2006) AliBaba: PubMed as a graph. *Bioinformatics*, **22**, 2444–2445.

Rocktäschel,T. *et al.* (2012) ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, **28**, 1633–1640.

Rodriguez-Esteban,R. *et al.* (2009) Identifying and classifying biomedical perturbations in text. *Nucleic Acids Res.*, **37**, 771–777.

Roy,S. *et al.* (2011) Latent semantic indexing of PubMed abstracts for identification of transcription factor candidates from microarray derived gene sets. *BMC Bioinformatics*, **12(Suppl. 10)**, S19.

Stam,C.J. (2005) Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. *Clin. Neurophysiol.*, **116**, 2266–2301.

Thagard,P. (2012) *The Cognitive Science Of Science: Explanation, Discovery, And Conceptual Change*. MIT Press, Cambridge, MA, USA.

Thomas,P. *et al.* (2012) GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res.*, **40**, W585–W591.