doi:10.1093/nar/gkn986

Identifying and classifying biomedical perturbations in text

Raul Rodriguez-Esteban*, Phoebe M. Roberts and Matthew E. Crawford

Pfizer Research Technology Center, 620 Memorial Dr., Cambridge, MA 02139, USA

Received September 11, 2008; Revised November 11, 2008; Accepted November 21, 2008

ABSTRACT

Molecular perturbations provide a powerful toolset for biomedical researchers to scrutinize the contributions of individual molecules in biological systems. Perturbations qualify the context of experimental results and, despite their diversity, share properties in different dimensions in ways that can be formalized. We propose a formal framework to describe and classify perturbations that allows accumulation of knowledge in order to inform the process of biomedical scientific experimentation and target analysis. We apply this framework to develop a novel algorithm for automatic detection and characterization of perturbations in text and show its relevance in the study of gene–phenotype associations and protein–protein interactions in diabetes and cancer. Analyzing perturbations introduces a novel view of the multivariate landscape of biological systems.

INTRODUCTION

In the early days of biological research, mutations that caused discernable phenotypes were the primary tool for understanding how a biological system worked—in the absence of a mutation a gene was invisible. Today, biologists are armed with a whole arsenal of tools to regulate gene, mRNA, and protein abundance and activity, thereby promoting the discovery of mechanisms and how a system gone awry can lead to disease (1). Among these are tools for suppressing the activity of a gene or gene product (e.g. site-directed mutagenesis, RNA interference, small molecule inhibitors) or enhancing activity (e.g. activating mutations or receptor agonist). Markedly different approaches can be used to perturb biological systems with similar effects. For instance, interfering with protein activity using small-molecule inhibitors should have a phenotype similar to reducing the abundance of the corresponding mRNA with anti-sense oligonucleotides (2). Likewise, similar responses are expected whether increases in intracellular protein concentration are achieved via an inducible promoter or by addition of recombinant protein (3). As such, perturbations form the core of understanding how biological systems work, how diseases arise, and how they can be treated. Any serious attempt to analyze a biological process starts by identification and characterization of perturbations that have been used in prior work. This task requires a framework that can be systematically applied and that is amenable to both manual and automatic means.

Currently, there is no established categorization that sufficiently represents the range of described experimental manipulations beyond high-level semantic and grammatical classifications (4,5) or description of techniques (6). For example, the closest concept we have found is ‘altered expression,’ defined as ‘altered expression level of a gene/protein’ (7). We believe that this concept is overly specific and fails to cover important phenomena, among others, changes in protein activity or gene mutations. We propose, instead, taking the existing concept of ‘perturbation’ and broadening it to comprise the range of terms used in text to indicate changes in the abundance or activity of DNA, RNA and proteins. Perturbations, in this new formulation, would refer to a collection of phenomena in a manner analogous to the way protein–protein interactions refer to biological phenomena of different type (e.g. bind, activate, inhibit). Since this proposition, like any other, needs to be tested for validity and utility, we have applied it to a case study involving gene–phenotype associations in disease and have developed a mining algorithm that detects the diverse forms in which perturbations appear in text. Therefore, we are introducing in this work both a new way to understand a crucial part of biology and a new text-mining method tailored to its extraction.

MATERIALS AND METHODS

We created three corpora that we named ‘design’, ‘test’ and ‘analysis’. As initial step, we created the design corpus to develop an analytical framework

*To whom correspondence should be addressed. Email: Raul.Rodriguez-Esteban@pfizer.com

Table 1. Examples of relevance and direction annotation

Relevance	Direction	Sentence	Explanation	PMID
Irrelevant		Different members of bcl-2 family may promote or inhibit apoptosis by synthesizing anti- and proapoptotic proteins	Members of a gene family, not a gene, are causing apoptosis phenotype	17404013
Irrelevant		The ATF6 and IRE1 pathways cooperatively caused apoptosis via induction of CHOP, activation of XBP1 and phosphorylation of JNK, and the PERK-eIF2α pathway counteracted the proapoptotic processes	Pathways, not genes, are causing apoptosis phenotype	17464326
Irrelevant		This result suggests that toxic products such as reactive oxygen species and aldehydes liberated by the action of polyamine oxidase on the acetylated polyamines formed by SSAT may enhance tumor development	Metabolites, not proteins that generate them, are causing tumor development phenotype	17675337
Relevant	Unknown	These results indicate that survivin is important for optimal development of bovine blastocysts and confirm that survivin expression suppresses apoptosis of pre-implantation embryos	'Survivin expression' does not indicate if there is a state change in amount	17075833
Relevant	Increase	Finally, we showed that after 7–9 days of incubation, MCH also inhibits proliferation of non-stimulated PBMC	'Incubation' indicates MCH was added exogenously	17537530
Relevant	Increase	Taken together, LDM induces apoptosis in a p53-dependent manner when given at low doses, but in a p53-independent manner when given at high doses	'Given at low doses' indicates LDM was added exogenously	17534142
Relevant	Decrease	Soluble Flt-1 abolished hypoxia/VEGF-induced hyperpermeability	Soluble forms of receptors create inactive complexes	17311300

for annotation. The purpose of this corpus was to identify challenges in the annotation process and to refine guidelines that would help the annotators in choosing their evaluations. Annotating perturbations requires at times thorough knowledge of experimental biology, which can only be captured and organized within a solid framework. Therefore we sought to perform a preliminary analysis on a test corpus to improve on subsequent annotations. The design corpus was not used for any other purpose. This corpus was limited to sentences that included disease-related gene–phenotype relationships. Using the semantic relationship nomenclature of Tsai *et al.* (8), we selected reports in which the 'agent' that deliberately performs an action is represented by a gene or protein, and the 'patient' that is the recipient of the action corresponds to disease phenotypes. The information we sought stands in contrast to associative relationships, such as elevated protein levels correlating with disease activity.

To create the design corpus, our initial query matched Medline sentences containing ordered triplets of a gene or protein name, a causative verb and a phenotype related to cancer or diabetes. Each member of the triplet was separated within the sentence by a maximum of three words. The retrieval was performed using Linguamatics I2E version 3.0 (Cambridge, UK). This software package has the ability to retrieve sentences from text that include word patterns established by user queries. The queries may include both syntactic and semantic constructs. Semantically, term classes can be defined combining external ontologie and adding term morphological variations and regular expression patterns. Syntactically, it recognizes part-of-speech and shallow parsing constructs such as noun phrases, verbs and prepositions. In terms of scope, queries can be confined to different document or text sections, including abstracts, titles or sentences. For example, a sentence-level query may comprise a term class protein, a list of verbs (I2E can automatically generate morphological variants) and a list of phenotype objects. The gene/protein thesaurus was internally developed and based on BioThesaurus (9). Forty verbs that signal causality (e.g. inhibition, stimulation) were compiled manually, as was a set of disease-related phenotypes relevant to cancer (e.g. tumorigenesis, vascularization) and diabetes (e.g. serum glucose levels, weight gain).

A set of 100 retrieved sentences and relationships were annotated by three PhD-level evaluators for relevance and direction of perturbation (see Table 1 for examples). 'Relevance' annotation was used to mark relationships that should be eliminated from the corpus for being irrelevant to the intent of the retrieval query, e.g. if the gene was not acting as an agent. 'Direction' indicated the type of perturbation (increased, decreased or unknown) relative to the starting state. For example, if a gene is added back to cells that carry a deletion in that gene to restore the wild-type state, it is noted as an increase, because the abundance of the gene was increased relative to the starting state. Sentences were annotated as 'unknown' if there was no stated perturbation, or direction could not be inferred at the sentence level. It is worth noting that although an unknown direction could be frequently resolved by reviewing the abstract or the full text of the article, we strictly limited our scope to evidence found at the sentence level.

While most sentences contained straightforward descriptions of perturbations (e.g. 'mutations in gene A' or 'protein B inhibition'), the broad range of perturbations in the literature, combined with the complex grammatical structures found in biomedical text, made some sentence–relationship pairs difficult to annotate consistently, mainly due to differences in knowledge of experimental descriptions and settings by the evaluators. When inter-annotator agreement proves to be challenging, other groups have adopted strategies to further improve the final gold standard annotation set (10). For this work, sentences deemed difficult to evaluate were set aside for later annotation by discussion and group consensus.

Therefore, the gold standard was comprised of annotations with three-way agreement from the individual assessment, plus the consensus annotations. When each evaluator's set of 'straightforward' independent annotations ($n = 237$ out of 300 annotations, 79%) was compared to the gold standard, agreement averaged 92.9%. Sentence–relationships for which there was no agreement or only two-way agreement were discussed and annotated by consensus (14%, $n = 14$). While overall inter-annotator agreement cannot be measured with the annotation procedure devised, the agreement metrics described are helpful as proxies to the level of ambiguity of the task. Pairs that are only evaluated by consensus differ from the rest largely in the knowledge required from the evaluator to elucidate the annotation, rather than in the factual ambiguity of the sentence. Hence, discussion and consensus assessment of pairs yields a better annotation set than individual annotation. An agreement of 92.9% can be considered high for the task. Only 7% ($n = 7$) of sentence–relationship pairs were marked irrelevant.

Following the same guidelines, we created the test corpus. Each of the three evaluators assessed different sets of 500 sentence–relationship pairs, 250 for diabetes and 250 for cancer. Sentences deemed not straightforward were left without annotation ($n = 126$, 8.4%) and later annotated by consensus. Only 6.3% of all relationships ($n = 95$) were considered irrelevant, demonstrating the high precision of the retrieval query. Overall, the procedure used to construct the corpus assured high quality to the 1405 sentence–relationship annotation. Genes were mapped to standard nomenclature using our protein/gene thesaurus. We measured the accuracy of this mapping using 250 sentences from the corpus. Accuracy was 89%.

We performed detection of increased, decreased or unknown perturbations using machine learning techniques. For that purpose, we constructed a vector of features for each relationship–sentence pair using the test corpus above. The vector was composed of different sets of features. One of them represented token presence as a binary vector of token weights w_i, $d = \{w_1, \ldots, w_{|T|}\}$, where T is the set of sentence tokens: a weight has value 1 if the token is present in the sentence and value 0 otherwise, an approach called set of words (SOW) (11). Tokens were created by stemming and tokenization of the sentence words. Hyphenated names were considered both as single and separate tokens in the SOW in order to capture affixes like 'anti-.' Since proximity to the gene name can be important in determining a token's role in describing a perturbation, the sentence was further divided in several sections relative to the gene name's position. Each section was characterized by a set of features using SOW. The sections considered were: n tokens before the gene name, where $n = \{5, 10, \text{all}\}$, and n tokens after the gene name, where $n = \{10, \text{all}\}$. We noted that many perturbation descriptions were adjacent or very close to the gene name (e.g. 'overexpression of p53'). Hence, we included a feature with the distance between the gene name and the beginning of the sentence (e.g. distance of 0 or 1 may indicate that the perturbation is unknown, e.g. 'TNF-α induces apoptosis...' or 'The p53 gene...'). We also created a set of features using a small perturbation ontology developed independently over a disease-agnostic set of retrieved sentences. If a member of the ontology was present in the sentence a feature was added with value 1, otherwise with value 0. All the feature sets described were integrated in a single feature vector for each sentence. An algorithm based on the principles of maximum entropy (12, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html) was trained using 80% of the test corpus, randomly sampled from the cancer and diabetes sets, and tested over the remaining 20% (results were averaged over 10 runs). The training allowed the algorithm to predict the presence of a perturbation and its direction. Performance measures were evaluated and compared favorably to an SVM algorithm (13).

To create the analysis corpus, we extended the scope of our methodology used to retrieve the test corpus by eliminating the disease-specific phenotype constraints in the retrieval query (hence, phenotypes were not included in the query). We applied the query to Medline diabetes abstracts after 1996 and retrieved 359 385 relationships related to different conditions and phenotypes. This output was run through the machine-learning algorithm to create a wide-scope, disease-agnostic set of 191 240 perturbation predictions. To create a diabetes-specific subset, only sentences from Medline abstracts containing the word diabetes in their MeSH descriptors were included.

RESULTS

There were significant differences both in technique and direction between the cancer and diabetes perturbations in the test corpus, with decreased perturbations more prevalent in cancer. Table 2 shows the performance of the

Table 2. Perturbation extraction performance

	Count	Precision (%)	Recall (%)	*F*-measure (%)
Perturbation	690	79.2	79.8	79.4
No ontology		76.4	79.6	77.6
Only ontology		77.2	59.5	67.0
Straightforward		83.3	78.9	81.0
Increased	436	75.3	71.1	72.9
No ontology		76.6	69.9	72.9
Only ontology		65.7	50.9	56.8
Straightforward		77.7	68.8	72.7
Decreased	254	79.0	65.1	71.2
No ontology		78.5	63.6	70.0
Only ontology		79.8	58.4	66.4
Straightforward		80.4	67.2	72.9

Precision, recall and macro-averaged *F*-measure were calculated using four different combinations of feature sets: 'full', 'no ontology', 'only ontology' and 'straightforward'. Baseline frequencies were 32.8% for increased perturbations and 18.1% for decreased perturbations.

Table 3. Ontology occurrence count in sentences from the test corpus, separated by diabetes and cancer phenotypes

Gene and protein perturbations	Cancer	Diabetes		Cancer	Diabetes
Total increasing modifications	266	586	Total decreasing modifications	267	110
General increasing modifications			General decreasing modifications		
Activation	66	12	Down-regulated, -ion	12	0
Administered, -ion	7	120	Inactivation	5	0
Dose, -age, -dependent	11	86	Inhibition	30	7
Ectopic	7	0	Repression	1	0
Enhanced … expression	10	0	Suppression	10	3
Exogenous	6	23	DNA decreasing modifications		
i.c.v.	0	11	Deficiency, -ent	15	12
i.p.	0	7	Deletion	3	2
Increased	26	18	Dominant-negative	3	0
Induction	8	2	Knockout	1	2
Infused, -ion	0	35	Loss	12	13
Injected, -ion	7	66	Mutant	0	0
Intracerebroventricular	0	51	Mutated, -ion	0	3
Intraperitoneal	1	16	mRNA decreasing modifications		
mg/kg,/kg	0	8	Anti(-)sense	9	3
Oral	6	4	Interference	5	1
Overexpressed, -ration	38	0	Interfering RNA	2	0
Peripherally	0	7	Knockdown	12	0
Recombinant	7	10	RNA interference	4	0
Restoration	2	0	RNAi	2	0
Subcutaneous	0	2	Short hairpin RNA	1	0
Systemic	1	9	Silenced	1	0
Treatment	24	22	siRNA	15	0
Up-regulation	8	0	Protein decreasing modifications		
DNA increasing modifications			Antagonist	7	28
Adenoviral	2	1	Anti-	25	4
Adenovirus	0	0	Antibody, -ies	11	3
Gain-of-function	2	0	Blockade, -ing	12	8
Gene delivery	1	3	Deactivation	1	0
Transgenic	3	2	Decoy	2	0
mRNA increasing modifications			Fc-	0	0
Inducible	2	0	Inhibitor	53	15
Protein increasing modifications			Inverse agonist	0	1
Activator	10	0	mAbs	1	0
Agonist	9	54	Neutralization	1	0
Analog/analogue	2	7	Reduced … activity	1	5
			Soluble	4	0
			Targeting	6	0

Note that total increased and decreased perturbations exceeds values in Table 2 since multiple perturbation terms may appear in a single sentence.

perturbation–detection algorithm built using different combinations of feature sets. Our algorithm detected perturbations with *F*-measure of 79.4%. Detection of increased and decreased perturbations had a lower *F*-measure, 72.9% and 71.2%, respectively. When we excluded the perturbation ontology in feature generation, results were only slightly lower, whereas when we exclusively used the perturbation ontology, results were much lower, notably due to reduced recall. ‘Straightforward’ relationships (91.6% of the total) were those that evaluators annotated without consultation with other annotators. These relationships were less challenging for humans, and the algorithm had better performance over this subset than overall. Due to absence of previous work in perturbation detection, these results cannot be compared, but they fall in ranges typical of other successful biomedical text mining tools.

To determine whether the disease impacted the frequency of perturbation types, we compared cancer and diabetes literature. The diabetes literature was significantly enriched in increased perturbations, whereas the cancer literature showed an even distribution between increased and decreased perturbations. In cancer, more perturbations were performed with antisense oligonucleotide ($n = 9$), antibody ($n = 16$) and RNA interference ($n = 26$) than in diabetes (antisense, $n = 3$; antibody, $n = 2$; RNAi, $n = 0$). Perturbations in diabetes were more frequently described as injections ($n = 132$) and/or administered by dose ($n = 57$) than in cancer (dose, $n = 3$; injection, $n = 6$). Perturbations in cancer were also more frequently described as being *in vivo* ($n = 12$) or *in vitro* ($n = 14$) than in diabetes (*in vivo*, $n = 5$; *in vitro*, $n = 0$). We examined the sentences for frequently occurring terms, grouped by their level of regulation (i.e. protein, mRNA or DNA), if known. Table 3 illustrates the wide variation in terms and affixes used in both diseases. Many of the terms related to dosing and routes of administration (e.g. administration, intracerebroventricular, injection) show a strong dominance in the diabetes literature compared to the cancer literature. Although these usually indicate an increasing perturbation, there can be exceptions, such as systemic delivery of an inhibitor.

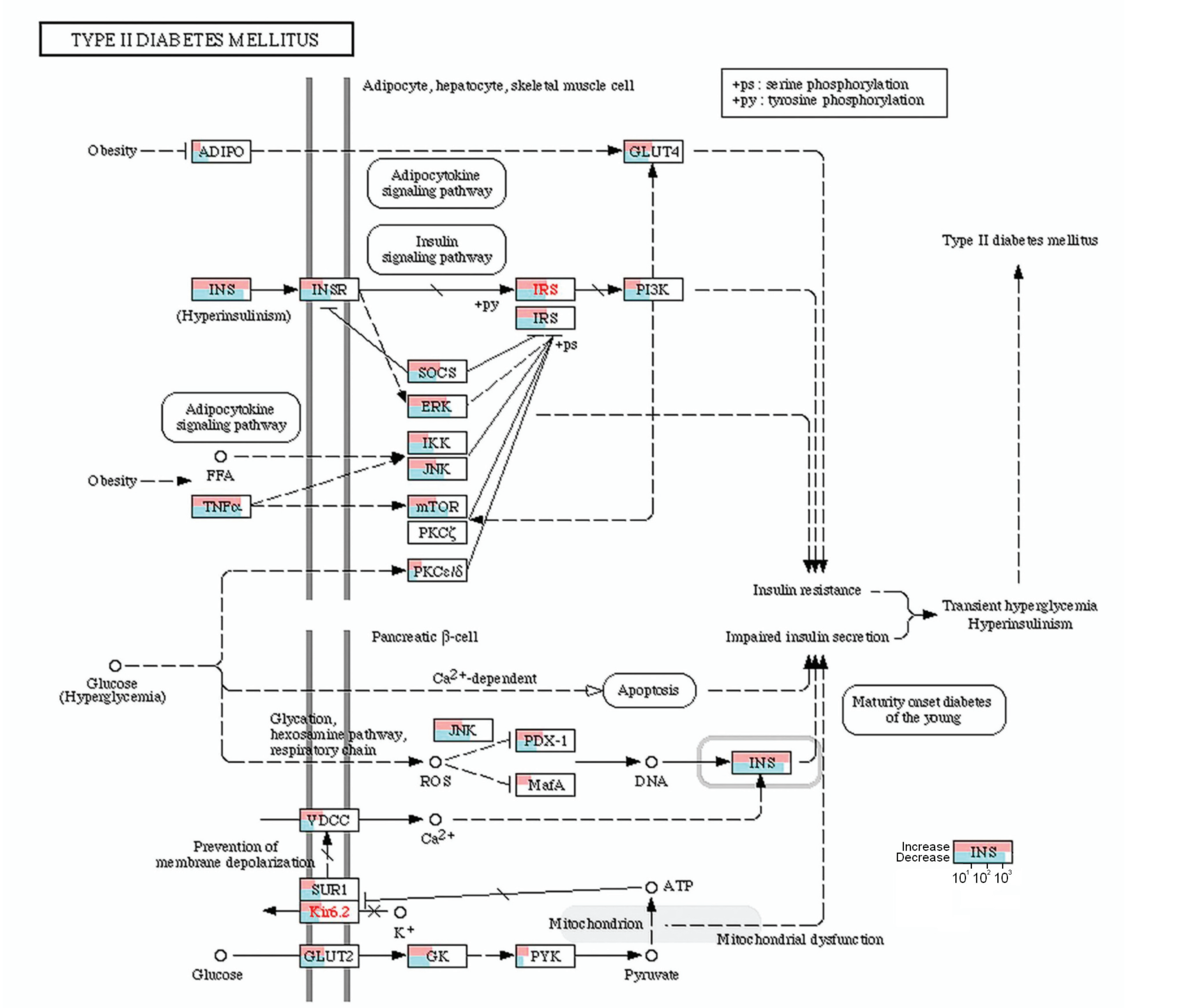
TYPE II DIABETES MELLITUS
Adipocyte, hepatocyte, skeletal muscle cell
+ps : serine phosphorylation
+py : tyrosine phosphorylation
Obesity
ADIPO
GLUT4
Adipocytokine signaling pathway
Insulin signaling pathway
Type II diabetes mellitus
INS
(Hyperinsulinism)
INSR
IRS
+py
IRS
PI3K
+ps
SOCS
ERK
Adipocytokine signaling pathway
IKK
JNK
FFA
Obesity
TNFα
mTOR
PKCζ
PKCε/δ
Insulin resistance
Transient hyperglycemia
Hyperinsulinism
Impaired insulin secretion
Pancreatic β-cell
Glucose
(Hyperglycemia)
Ca2+-dependent
Apoptosis
Maturity onset diabetes of the young
Glycation, hexosamine pathway, respiratory chain
JNK
PDX-1
MafA
ROS
DNA
INS
VDCC
Ca2+
Prevention of membrane depolarization
Increase
Decrease
INS
10¹ 10² 10³
SUR1
Kir6.2
K+
ATP
Mitochondrion
Mitochondrial dysfunction
Glucose
GLUT2
GK
PYK
Pyruvate

Figure 1. Perturbations extracted from genes involved in the Diabetes Mellitus Type II pathway in the Kyoto Encyclopedia of Genes and Genomes (21). Differences in perturbation direction are large for some members of the pathway, note that the scale bar is logarithmic.

The phenotypes in this study were selected based on prevalence in the literature without regard to their use *in vivo* or *in vitro*. For instance, in diabetes a change in blood glucose levels signals a change in disease state—clearly a phenotype monitored following perturbation *in vivo*. Likewise, in cancer a change in the number of metastases is exclusively described in *in vivo* systems. In contrast, insulin sensitivity can be used to describe both *in vivo* and *in vitro* systems. For cancer, cell proliferation and apoptosis rates are also described both *in vivo* and *in vitro*.

If our hypothesis that perturbed genes and proteins form the underpinnings of disease mechanisms, these entities should be well represented in pathway diagrams and among drug candidates. We applied our trained algorithm to a set of 14 345 sentence–relationship facts for genes from our analysis corpus belonging to a diabetes pathway (Figure 1). Predictably, our algorithm found a strong correlation between the number of Medline abstracts in which a gene is mentioned and the number of times it is described as perturbed ($r^2 = 0.70$). The results in Figure 1 show the difference in intensity and modality in which each gene is described or pursued experimentally. Some genes were typically more increased or decreased, often reflecting their roles in therapeutics. Examples of significantly ($P < 10^{-6}$) increased were insulin, interleukin 2 or parathyroid hormone. Among the decreased were such genes as epidermal growth factor receptor; caspase 8 and plasminogen activator, urokinase receptor.

We compared detected perturbation counts from the analysis corpus against the gene–disease associations included in the OMIM Morbid Map (OMIM online reference, http://www.ncbi.nlm.nih.gov/omim/). The average abstract mention and gene perturbation counts were higher for genes with MorbidMap associations (*t*-test, $P < 10^{-4}$). This was consistent with our expectation that genes associated to disease would be the subject of deeper study. However, genes that had been perturbed numerous times were not necessarily linked to disease. For example, out of the 100 most perturbed genes, 54 were not linked to disease in OMIM MorbidMap, including at the top such genes as jun oncogene, interleukin 4, fibroblast growth factor 2 (basic), colony stimulating factor 2 (granulocyte–macrophage), colony stimulating factor 3 (granulocyte) and mitogen-activated protein kinase 3.

DISCUSSION

Perturbations are relevant for areas like the study of gene–disease associations, protein–protein interactions (PPI) or gene regulatory networks. Gene–disease association extraction studies have largely focused on simply detecting associations rather than characterizing them (7,14). PPI and gene regulation extraction studies have side-stepped perturbation types, without considering anything further than the experimental technique (15,16). We note that PPI relationships are frequently devoid of perturbations and the focus is on how a PPI was detected, which does not necessarily involve a causative relationship. The Proteomics Standards Initiative – Molecular Interactions (PSI-MI) ontology (6) is a comprehensive effort to describe molecular interactions. This ontology, while including a detailed experimental preparation section, lacks expressivity in describing perturbations generically. Similarly to Bundschus *et al.* (7), it includes a section on expression level with entries ‘under expressed’, ‘over expressed’ and ‘physiological’.

The lack of a general framework for recognizing, characterizing and classifying perturbations is surprising when one considers their importance in phenomena encountered experimentally. Researchers with interest in characterizing previous and current perturbation work on a biological system face the challenge of a naturalist trying to deal with animal species without a Linnaean taxonomy. This is reflected in the methodological landscape that was set in the early literature in the fields of biomedical ontologies and text mining. For example, the comprehensive text mining tool MedLEE (17) only considered the ‘state’ of a gene or protein, where the state has an adjectival role such as ‘mutated’ in the phrase ‘mutated X’. Perturbation descriptions, however, should be considered carefully. Observe the differences between the sentences (i) ‘X activates Y.’ and (ii) ‘Inhibition of X activates Y’. From the point of view of classic PPI extraction both relationships are equivalent and can be represented with the triplet *X activate Y*. The perturbation in sentence (ii), however, signals that it is likely that protein X is inhibiting protein Y instead. We have called this phenomenon ‘reversal’. Given the results of the present assessment, a review of the relationship data available should be considered under this model.

We have focused our methodology in gene–phenotype associations in disease but the principles shown are applicable to other well-known areas, such as PPI, as well as less explored ones such as identification of biomarkers or cellular processes. A perturbation taxonomy, like the one described in Table 4, could capture the different dimensions that may be of interest to the inquiring scientist. This taxonomy has four annotation types: relevance, direction, molecule and effect. Relevance annotation marks relationships that are irrelevant to the intent of the retrieval query. Direction distinguishes between perturbations that represent an increase or a decrease over starting levels. Unknown direction annotation is intended for perturbations whose direction cannot be inferred at the sentence level. Molecule annotation characterizes the type of molecule being primarily affected by the perturbation: gene, RNA or protein. Expression annotation is used for a change of expression level without clarifying whether the change is in RNA or protein. Effect annotation differentiates between changes in activity or function and changes in abundance. The following examples illustrate these annotations: A gene mutation is a decrease in gene activity, where the function/activity of a gene is specifically understood as making a wild-type transcript. Exogenous addition of a gene via viral transfection, plasmid transformation, etc., is an increase in gene abundance. A gene duplication is an increase in gene abundance while a knockout is a decrease in gene abundance. Dominant negative is a mutation in a gene, which indicates that, compared to wild-type, it has defective function. Silencing, knock-down and antisense all apply to a decrease in RNA abundance. An antibody blocking a protein decreases the protein activity or function. Interfering with a protein binding another protein is a decrease on a protein’s function or activity. Treatment, incubation, recombination, or synthetic refer to exogenous addition of protein.

Table 4. Proposed annotation guidelines

Relevance	Direction	Molecule	Effect
Relevant	Increase	Gene	Activity
Irrelevant	Decrease	RNA	Abundance
	Unknown	Protein expression	

Perturbations evolve, notably as new techniques are developed and targets are identified. We expect perturbations to be subject to trends and popularity variations similar to those in other aspects of biomedicine (18–20).

ACKNOWLEDGEMENTS

We would like to thank Michael L. McGlashen for his continued support to the work that led to this manuscript.

FUNDING

Funding for open access charge: Pfizer Systems Biology.

Conflict of interest statement. None declared.

REFERENCES

1. Manipulating Proteins, DNA and RNA. In Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P., (2002) *Molecular Biology of the Cell*. 4th edn. Garland, New York.
2. Evans,R., Naber,C., Steffler,T., Checkland,T., Keats,J., Maxwell,C., Perry,T., Chau,H., Belch,A., Pilarski,L. *et al.* (2008) Aurora A kinase RNAi and small molecule inhibition of Aurora kinases with VE-465 induce apoptotic death in multiple myeloma cells. *Leuk. Lymphoma*, **49**, 559–569.
3. Providence,K.M., Higgins,S.P., Mullen,A., Battista,A., Samarakoon,R., Higgins,C.E., Wilkins-Port,C.E. and Higgins,P.J. (2008) SERPINE1 (PAI-1) is deposited into keratinocyte migration "trails" and required for optimal monolayer wound repair. *Arch. Dermatol. Res.*, **300**, 303–310.
4. Friedman,C., Kra,P. and Rzhetsky,A. (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J. Biomed. Inform.*, **35**, 222–235.
5. Pyysalo,S., Ginter,F., Heimonen,J., Björne,J., Boberg,J., Järvinen,J. and Salakoski,T. (2007) BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, **8**, 50.
6. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
7. Bundschus,M., Dejori,M., Stetter,M., Tresp,V. and Kriegel,H.P. (2008) Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, **9**, 207.
8. Tsai,R.T., Chou,W.C., Su,Y.S., Lin,Y.C., Sung,C.L., Dai,H.J., Yeh,I.T., Ku,W., Sung,T.Y. and Hsu,W.L. (2007) BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, **8**, 325.
9. Liu,H., Hu,Z.Z., Zhang,J. and Wu,C. (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.
10. Colosimo,M.E., Morgan,A.A., Yeh,A.S., Colombe,J.B. and Hirschman,L. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, **6(Suppl. 1)**, S12.
11. Sebastiani,F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys*, **34**, 1–47.
12. Berger,A.L., Della Pietra,V.J. and Della Pietra,S.A. (1996) A maximum entropy approach to natural language processing. *Comput. Linguist.*, **22**, 39–71.
13. Joachims,T. (1998) Making large-scale support vector machine learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, p. 392.
14. Chun,H.W., Tsuruoka,Y., Kim,J.D., Shiba,R., Nagata,N., Hishiki,T. and Tsujii,J. (2006) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics*, **7(Suppl. 3)**, S4.
15. Rzhetsky,A., Koike,T., Kalachikov,S., Gomez,S.M., Krauthammer,M., Kaplan,S.H., Kra,P., Russo,J.J. and Friedman,C. (2000) A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics*, **16**, 1120–1128.
16. Beisswanger,E., Lee,V., Kim,J.J., Rebholz-Schuhmann,D., Splendiani,A., Dameron,O., Schulz,S. and Hahn,U. (2008) Gene Regulation Ontology (GRO): design principles and use cases. *Stud. Health Technol. Inform.*, **136**, 9–14.
17. Friedman,C., Alderson,P.O., Austin,J.H., Cimino,J.J. and Johnson,S.B. (1994) A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.*, **1**, 161–174.
18. Cokol,M., Iossifov,I., Weinreb,C. and Rzhetsky,A. (2005) Emergent behavior of growing knowledge about molecular interactions. *Nat. Biotechnol.*, **23**, 1243–1247.
19. Pfeiffer,T. and Hoffmann,R. (2007) Temporal patterns of genes in scientific publications. *Proc. Natl Acad. Sci. USA*, **104**, 12052–12056.
20. Cokol,M. and Rodriguez-Esteban,R. (2008) Visualizing evolution and impact of biomedical fields. *J. Biomed. Inform.*, **41**, 1050–1052.
21. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.