Molecular Biomarkers & Diagnosis

Trugenberger and Peregrim, J Mol Biomark Diagn 2013, S3
http://dx.doi.org/10.4172/2155-9929.S3-004

Research Article Open Access

Discovery of Novel Biomarkers by Text Mining: A New Avenue for Drug Research?

Carlo A Trugenberger[1*†] and David Peregrim[2‡]

[1]*InfoCodex AG, Semantic Technologies, Bahnhofstrasse 50, Buchs (SG) CH-9470, Switzerland*
[2]*Merck Research Laboratories, 126 East Lincoln Avenue, Rahway, NJ 07065, USA*
[†]*Primary author*
[‡]*Secondary contributor*

Abstract

Data are paramount to modern targeted drug design. Precious revelations obtained by applying data mining and computational chemistry on large molecular databases, innovative at one time, are now everyday procedures for therapy identification. However, there is an even larger source of valuable information available that can potentially be tapped for discoveries: repositories constituted by research documents.

While numerical methods for the analysis of structured data like those in genomics and proteomics databases are well developed and standard toolboxes are easily available, knowledge discovery from unstructured data in text documents is still considered the "Holy Grail" of text mining and no stable methodology has yet emerged from the scant few known attempts.

Here we review a recent pilot experiment to discover novel biomarkers and phenotypes for diabetes and obesity by self-organized text mining of about 120,000 PubMed abstracts, public clinical trial summaries, and internal Merck research documents by the InfoCodex semantic engine. Retrieval of known entities missed by other traditional approaches could be demonstrated and the InfoCodex semantic engine was shown to discover new diabetes and obesity biomarkers and phenotypes, although noticeable noise (uninteresting or obvious terms) was generated.

The reported text mining approach to biomarker discovery shows much promise and has the potential to be developed into a new avenue for pharmaceutical research, especially to shorten time-to-market of novel drugs, or speed up early recognition of dead ends and adverse reactions.

Keywords: Biomarker discovery; Text mining; Semantic technologies; Biomedical ontologies

Abbreviations: ADME: Absorption/Distribution/Metabolism/Excretion; CL: Confidence Level; CRO: Contract Research Organization; CUI: Concept Unique Identifier; D&O: Diabetes & Obesity; DM1: Diabetes Mellitus Type 1; DM2: Diabetes Mellitus Type 2; DI: Diabetes Insipidus; GO: Gene Ontology; ILD: Infocodex Linguistic Database; IR: Information Retrieval; NLP: Natural Language Processing; NMR: Nuclear Magnetic Resonance; OMIM: Online Mendelian Inheritance In Man; P3: Merck Internal Research Documents Database; RO: Related Other; RN: Related Narrow; SME: Subject Matter Expert; SOM: Self-Organizing Map; TGI: Target-Gene Information

Background

Information extraction and knowledge discovery in research papers

Pharmaceutical research is undergoing a profound change. The deluge of molecular data and the advent of computational approaches to analyze them have revolutionized the traditional process of discovering drugs by happenstance in natural products or synthetizing and screening large libraries of small molecule compounds. Today, computational methods permeate so many aspects of pharmaceutical research that one can say that drugs are "designed" rather than "discovered" [1,2].

Molecular data found in genomics and proteomics databases are typically structured (well organized in databases or hierarchical schemes) data. Numerical methods to analyze this type of data have a long history and are well developed [2]. Unfortunately, structured data constitute only the minority of the deluge of data the world is accumulating: it has been estimated [3] that 85% of the data stored on the world's computers are unstructured, with no identifiable organization, free text being the most common example.

This is no different in the pharmaceutical industry. While the bulk of the computational effort goes into crunching structured molecular data, there is another, even larger source of valuable information that can potentially be tapped for discoveries: repositories constituted by research documents. One of the best known of these repositories, PubMed, contains already more than 20 millions citations and these are growing at a once inconceivable rate of almost 2 papers/second [4].

The value of the information in these repositories of research is huge. Each paper by itself constitutes typically a much focused study on one particular biomedical subject that can be easily comprehended by other experts in the same field. It is to be expected there are also far-reaching correlations between the results of different papers or different groups of papers. Uncovering such hidden correlations by hand borders on the impossible since, first, the quantity of such papers accumulated by now are far beyond the reach of human analysis and,

***Corresponding author:** Carlo A Trugenberger, InfoCodex AG, Semantic Technologies, Bahnhofstrasse 50, Buchs (SG) CH-9470, Switzerland, E-mail: c.trugenberger@infocodex.com

Received May 02, 2013; **Accepted** June 18, 2013; **Published** June 20, 2013

Citation: Trugenberger CA, Peregrim D (2013) Discovery of Novel Biomarkers by Text Mining: A New Avenue for Drug Research? J Mol Biomark Diagn S3: 004. doi:10.4172/2155-9929.S3-004

secondly, the expertise to understand papers in different areas of research is very hard to find in the same individual in today's era of ever increasing specialization. The potential competitive advantage for the first companies to succeed in the task of discovering new scientific knowledge this way is considerable, both in speeding up research and in cutting costs.

Unfortunately, the available techniques for the analysis of unstructured data are far less developed than their arranged and classified numerical counterparts. The main (and best known) automated manipulation of unstructured data today is restricted to "information extraction" in both its classical "search" form based on keywords or in its more advanced version relying on natural language processing (NLP) [5,6]. Information extraction aims to recover knowledge that is explicitly stated in natural language documents. This is the domain of NLP systems that typically analyze documents sentence by sentence. By definition, this procedure can recover only known information, information that has been written down by some human expert in a document.

On the contrary, "knowledge discovery", the discovery of new facts by unveiling hidden associations is still considered the "Holy Grail" of text mining [5] and is a much more difficult task, in its infancy in the innovation curve, and without established approaches. A disruptive mechanism is needed at this time to exploit the wealth of hidden information in large research repositories. To meet this demand one must go beyond NLP to systems that by combining semantics and machine intelligence, are capable of analyzing document collections as a whole, and are thereby positioned to uncover possible associative, semantically unspecified relationships. The InfoCodex semantic engine is a tool designed specifically for this discovery task.

In order to explore the power of semantic machine intelligence for the screening of a collection of research documents in search of unknown/novel information relevant to early-stage drug candidate discovery and development, Merck, in collaboration with Thomson Reuters, devised a pilot experiment in which the InfoCodex semantic engine was used for the specific task to discover unknown/novel biomarkers and phenotypes for diabetes and/or obesity (D&O) by semantic machine analysis of diverse and numerous biomedical research texts [7].

Biomarkers and phenotypes

The pilot experiment was focused on biomarkers and phenotypes since these play a paramount role in modern medicine. Drugs of the future will be targeted to populations and groups of individuals with common biological characteristics predictive of drug efficacy and/or toxicity. This practice is called "individualized medicine" or "personalized medicine" [1]. The revealing features are called "biomarkers" and "phenotypes".

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. In other words, a biomarker is any biological or biochemical entity or signal that is predictive, prognostic, or indicative of another entity, in this case, diabetes and/or obesity.

A phenotype is an anatomical, physiological and behavioral characteristic observed as an identifiable structure or functional attribute of an organism. Phenotypes are important because phenotype-specific proteins are relevant targets in basic pharmaceutical research.

Biomarkers and phenotypes constitute one of the "hot threads" of diagnostic and drug development in pharmaceutical and biomedical research, with applications in early disease identification, identification of potential drug targets, prediction of the response of patients to medications, help in accelerating clinical trials and personalized medicine. The biomarker market generated $13.6 billion in 2011 and is expected to grow to $25 billion by 2016 [8].

Methods

The task of the experiment

The object of the experiment was for the InfoCodex semantic engine to discover unknown/novel biomarkers and phenotypes for diabetes and/or obesity (D&O) by text mining a diverse and sizable corpus of unstructured, free text biomedical research documents constituted by:

- PubMed [9] abstracts with titles: the 115,273 most recent documents (since 1/1/1998) retrieved by the query *diabetes OR obesity OR X* where *X* is a set of 27 known or suspected D&O biomarkers known to Merck and connected by Boolean OR's (i.e., *X* stands for *5HT2c OR AMPK OR DGAT1 OR FABP_4_aP2 OR FTO OR …*). The 27 biomarkers were supplied by the Diabetes and Obesity Merck franchise and consisted of, predominantly, genes relevant to those disorders.
- Clinical Trials [10] summaries: the 8,960 most recent summaries (since 1/1/2007) retrieved by the query *diabetes OR obesity*. (Adding the 27 Merck D&O biomarkers to the query did not result in any additional hits.)
- Internal Merck research documents, about one page in length: 500 documents. Merck internal research documents refer to a database of full summaries, figures, tables, conclusions, and other key molecular profiling project information predominantly in the fields of atherosclerosis, cardiovascular, bone, respiratory, immunology, endocrine, diabetes, obesity, and oncology.

The output D&O related biomarkers and phenotypes proposed by the machine were then compared with Merck internal and external vocabularies/databases including UMLS [11], GenBank [12], Gene Ontology [13], OMIM [14], and the Thomson Reuters [15] D&O biomarker databases.

By design, the experiment was handled strictly as a "blind experiment": no expert input about D&O biomarkers/phenotypes was provided and no feedback from preliminary results was used to improve the machine-generated results.

The InfoCodex semantic engine

InfoCodex is a text analysis technology designed for the unsupervised semantic clustering and matching of multi-lingual documents [16]. It is based on a combination of a universal knowledge repository (the InfoCodex Linguistic Database, ILD), statistical analysis and information theory [17], and self-organizing neural networks (SOM) [18].

InfoCodex linguistic database [ILD]: The ILD contains multi-lingual entries (words/phrases), each characterized by:

- its type (noun, verb, adjective, adverb/pronoun, name)
- its language (en, de, fr, it, es)
- its significance rank from 0 (meaningless glue word) to 4 (very significant and unique)

- a hash code for the accelerated recognition of collocated expressions

The words/phrases with almost the same meaning are collected into cross-lingual synonym groups (microscopic semantic clouds) and systematically linked to a hypernym (taxon) in universal 7-level taxonomy (simplified ontology restricted to hierarchical relations).

With its 3.5 million classified entries, the ILD corresponds to a very large multi-lingual thesaurus (for comparison, the *Historical Thesaurus of the English Oxford Dictionary*, often considered the largest in the world, has 920,000 entries). The content and the semantic structure of the ILD are largely based on WordNet [19], combined with some 100 other well established knowledge sources.

Text mining and content analysis: The words/phrases found in a document are matched with the entries in ILD, providing a cross-language content recognition. The taxons most often matched by a document represent the document's main topics. Using statistical methods and information theoretical principles, such as entropies of individual words, a 100-dimensional content space are constructed that can depict the document characteristics in an optimal way. The documents are then projected into this content space, resulting in 100-dimensional vectors characterizing the individual documents together with a generated set of the most relevant synonym groups.

Categorization of a document collection (Kohonen Map): The fully automatic categorization is achieved by applying the neural network technique of Kohonen [18], which creates a thematic landscape according to and optimized for the thematic volume of the entire document collection. Prior to starting the unsupervised learning procedure, a coarse group rebalancing technique is used to construct a reliable initial guess for the SOM. This is a generalization of coarse mesh rebalancing [20] to general iterative procedures, with no reference to spatial equation as in the original application to neutron diffusion and general transport theory in finite element analysis. This procedure considerably accelerates the iteration process and minimizes the risk of getting stuck in a sub-optimal configuration.

For the comparison of the content of different documents with each other and with queries, a similarity measure is used which is composed of the scalar product of the document vectors in the 100-dimensional content space, the reciprocal Kullback–Leibler distance [21] from the main topics, and the weighted score-sum of common synonyms, common hypernyms and common nodes on higher taxonomy levels.

As a result of the semantic SOM algorithm, a document collection is grouped into a two-dimensional array of neurons called an information map. Each neuron corresponds to a semantic class; i.e., documents assigned to the same class are semantically similar. The classes are arranged in such a way that the thematically similar classes are nearby (Figure 1).

The described InfoCodex algorithm is able to categorize unstructured information. In a recent benchmark, testing the classification of multi-lingual, "noisy" Web pages, InfoCodex reached the high clustering accuracy score F1 = 88% [22]. Moreover, it extracts relevant facts not only from single documents at hand, but it considers document collections as a whole and identifies dispersed and seemingly unrelated facts and relationships like assembling the scattered pieces of a puzzle.

Discovering biomarkers/phenotypes with InfoCodex

Four steps were involved in the procedure:

1. Create reference models: teaching the software the essential meaning of "what is a biomarker or a phenotype for D&O."
2. Determine the meaning of unknown terms (not part of the current ILD) in the document collection by semantic inference using the internal ILD knowledge base.
3. Identify candidates for D&O biomarkers/phenotypes by comparing the subset of documents containing the candidates with the reference models established in Step 1.
4. Compute confidence levels for the identified candidates.

Step 1: Reference models: In order to solve the task of the experiment, the InfoCodex semantic engine had to "comprehend" the meaning of "biomarker/phenotype for D&O". To this end, a training set of known biomarkers and phenotypes for D&O was determined by naïve (no input by human subject matter experts [SME] was provided) literature search via the autonomous InfoCodex spider agents. Of course, this search was independent of the 27 D&O biomarkers/phenotypes used by Merck to assemble the documents base via the PubMed query. This resulted in a list of 224 reference D&O biomarkers/phenotypes.

Four subsets of documents were identified containing these reference terms and the terms "diabetes" or "obesity" and "biomarker" or "phenotype" (2x2 matrix). Each of these subsets was then clustered into 5–6 subgroups such that the documents in each subgroup were semantically similar to each other using agglomerative hierarchical clustering [23].

For each of the 5–6 sub-clusters, an InfoCodex reference feature vector was then determined for later comparison. This reference feature vectors represent mathematical models formulated on the ILD of what, e.g., "biomarker of diabetes" means.

Step 2: determination of the meaning of unknown terms: While the ILD contained at the time of the experiment about 20,000 genes and proteins (up to around 100'000 presently), it was not guaranteed to identify all possibly relevant candidates by a simple database look-up. A procedure to infer the meaning of unknown terms from this "hard-wired" knowledge and for synonym analysis [24] had to be devised.

To describe the meaning of an unknown term, a hypernym (superordinate term) is constructed, which corresponds to a known taxon (node) in the taxonomy tree of the ILD. For example, the term "endocannabinoid" was not part of the ILD and, therefore, its meaning was unknown. However, if a procedure can assign the known taxon "receptor" as its most likely hypernym, the unknown term receives a meaning in the sense "is a".

The taxonomic hypernyms are inferred by information-theoretic algorithms which analyze the co-occurrences of the unknown terms with entries in the ILD and score the aggregated hypernyms of these cross-terms to select the most probable one. The top scoring hypernym of the cross-terms is selected as the "constructed hypernym" for the unknown term (Table 1). In an analogous way InfoCodex determines also associated descriptors for each unknown term: these have to be understood as horizontal "has to do with" relations, as opposed to vertical "is a" hypernym relations (Table 1).

InfoCodex computed meanings of some unknown terms from the experimental PubMed collection.

The meaning of unknown terms is estimated fully automatically; i.e., no human interventions were necessary and no context-specific

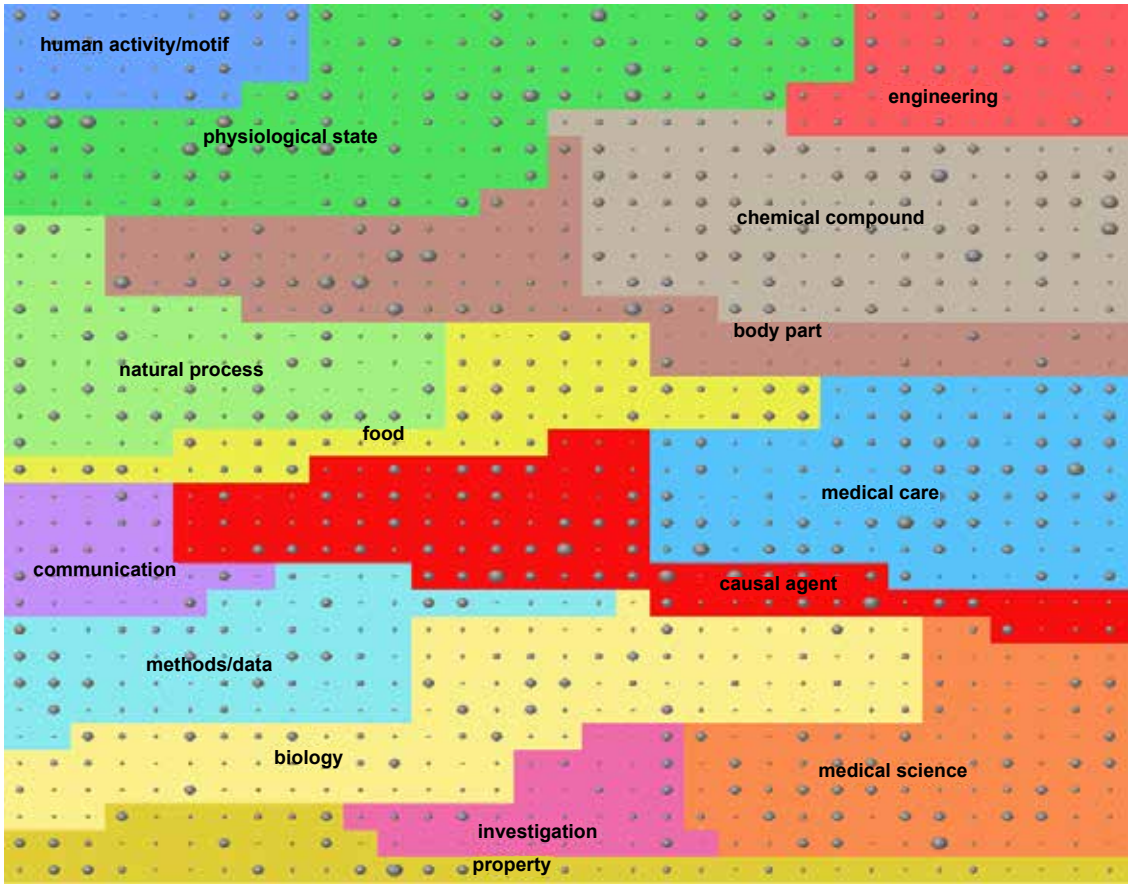

Figure 1: InfoCodex information map. InfoCodex information map obtained for the approximately 115,000 documents of the PubMed repository used for the present experiment. The size of the dots in the center of each class indicate the number of documents assigned to it.

Unknown Term	*Constructed Hypernym*	*Associated Descriptor 1*
Nn1250	clinical study	insulineglargine
Tolterodine	cavity	overactive bladder
Ranibizumab	drug	macular edema
Nn5401	clinical study	insulin aspart
Duloxetine	antidepressant	personal physician
Endocannabinoid	receptor	enzyme
Becaplermin	pathology	ulcer
Candesartan	cardiovascular disease	high blood pressure
Srt2104	medicine	placebo
Olmesartan	cardiovascular medicine	amlodipine
Hctz	diuretic drug	hydrochlorothiazide
Eslicarbazepine	anti nervous	Zebinix
Zonisamide	anti nervous	Topiramate Capsules
Mk0431	antidiabetic	sitagliptin
Ziprasidone	tranquilizer	major tranquilizer
Psicofarmcolagia	motivation	incentive
Medoxomil	cardiovascular medicine	amlodipine

Table 1: InfoCodex computed meanings.

vocabularies had to be provided as in most related approaches [6]. The meaning had to be inferred by the semantic engine only based on machine intelligence and its internal generic knowledge base, and this automatism is one of the main innovations of the presented approach. Some of the estimated hypernyms are completely correct: "Hctz" is a diuretic drug and is associated to "hydrochlorothiazide" (actually a synonym). "Duloxetine" is indeed an antidepressant, and the associated descriptor "personal physician" expresses the fact that the contact with the physician plays an important role in ("is related to") antidepressant usage. Clearly, not all inferred semantic relations are of the same quality.

Step 3: generating a list of potential biomarkers and phenotypes:

Most of the reference biomarkers and phenotypes found in the literature (see Step 1) were linked to one of the following nodes of the ILD:

- *Genes* (including the subnodes "nucleic acids" and "regulatory genes")
- *Proteins* (including the subnodes "enzymes", "transferase", "hydrolase", "antibodies", "simple proteins")
- *Causal agents* (including subnodes such as "anesthetics", "diuretic drugs", "digestive agents")
- *Hormones*
- *Phenotypes*
- *Metabolic disorders*
- *Diabetes*
- *Obesity*
- *Symptoms* (including the subnode "syndromes")

Each of the terms appearing in the experimental document base that point to one of these taxonomy nodes, whether via hypernyms given in the ILD for known terms or via constructed hypernyms for unknown terms, are considered as initial potential biomarker/phenotype candidates.

These are, then further assessed by forming document subsets of the experimental document base containing a synonym of one particular candidate in combination with synonyms of "diabetes" or "obesity" respectively.

Each document is characterized by a feature vector, which is determined by the InfoCodex self-organized analysis. The document subsets corresponding to one particular initial candidate are compared with the previously derived reference models for D&O biomarkers/phenotypes by computing the semantic distances of all feature vectors of the subset to the feature vectors of the reference models. A term qualifies as a final candidate for a D&O biomarker or phenotype if most of these semantic similarity deviations from one of the corresponding reference clusters are below a certain threshold.

Step 4: confidence levels: Not all the biomarker/phenotype candidates established this way have the same probability of being relevant. In order to rank the final candidates established in Step 3 an empirical score was devised, representing the confidence level of each term. This confidence measure is based on the average semantic deviation of the feature vectors assigned to the candidate from the feature vector of the corresponding reference model and additional information-theoretic measures.

Evaluation of results

The standard evaluation metrics in pattern recognition, information retrieval and classification are *precision* and *recall*, defined as follows for the case at hand:

- *Precision*: % of InfoCodex outputs matched by benchmark biomarkers and phenotypes.
- *Recall*: % of benchmark biomarkers and phenotypes matched by InfoCodex outputs.

Unfortunately, in the present experiment these measures cannot be used as proper evaluation metrics. On one side, *recall* would only be an accurate measure for the retrieval power if the reference vocabularies were established on exactly the same document corpus used in the experiment. This is not the case, since comprehensive biomarker repositories, such as Thomson Reuters' e.g., are based on amuch broader basis than the 120,000 PubMed abstracts used as a document sample in the current experiment. On the other side, *precision* is a relevant measure only if the benchmark is a comprehensive list of all possible items that could be retrieved. This is also not the case: new biomarkers/phenotypes, not recorded in any benchmark are the main issue of the present experiment. On the contrary, the *novelty* component of any biomarker database is zero by definition, which would lead to a strongly reduced *precision* in the assessment of the InfoCodex results (since *precision* is equal to 100% - *novelty*).

In the present experiment, the *human assessment* of valuable and irrelevant novel candidates is thus the most crucial evaluation. The objective of the experiment was not a statistically significant certification of a specific biomarker, but it was a proof-of-concept for the automatic discovery of novel biomarkers/phenotypes. Nevertheless, precision and recall measures were estimated to provide at least a qualitative indication of emerging trends.

Reference vocabularies/databases

For this qualitative evaluation, the InfoCodex-computed D&O biomarker and phenotype candidates were compared with Merck internal and external benchmark vocabularies/databases including UMLS [11], GenBank [12], Gene Ontology [13], OMIM [14], and Thomson Reuters [15], D&O biomarker databases.

Merck-internal vocabularies: I2E: As stressed above, a really meaningful recall assessment requires a reference list based on the exact same document pool used for the experiment. This is clearly not the case for the available standard databases described below. In order to obtain a rough estimate of such a reference list we used the Merck implementation of Linguamatics I2E [25], a text mining tool, to extract relevant class1-relation-class2 triples found within sentences in the experimental PubMed collection. This NLP tool provided a query-specific method to convert unstructured sentences mentioning biomarkers/phenotypes into a structured term list. It also serves as an example of the typical use of NLP tools as an aid in information extraction of known, lexicalized named entities, for comparison with the associative discovery approach of InfoCodex.

I2E-raw: I2E was used to extract relevant class1-relation-class2 triples found within sentences in the experimental PubMed collection. For biomarkers, class2 was defined as "diabetes" or "obesity" (note that no synonyms or hyponyms were used) and the relation as "biomarker" or any of its synonymous, lexical, or hyponymic variants according to the Linguamatics ontology. Class1 thus encompassed the I2E-extracted biomarkers. The result was 1,339 such triples; these triples could be de-duplicated, frequency-weighted, and reduced to 788 unique biomarkers for diabetes and 242 for obesity. The same procedure for phenotypes yielded 6,691 unique phenotypes for diabetes and obesity together.

I2E-normalized: The raw I2E phenotype output was normalized by one of Merck's Linguamatics consultants using automated mapping of the class2 values to UMLS controlled vocabulary terms, resulting in 12,015 unique triples, or 1,520 unique phenotypes for diabetes and obesity together.

I2E-manual: We manually extracted a curated version from the I2E-extracted PubMed sentences. This yielded 3,800 biomarker

triples; after de-duplication and synonym/variant conflation, 823 unique biomarkers for diabetes and 315 for obesity. It also yielded 11,365 phenotype triples; after de-duplication and synonym/variant conflation, 4,780 unique phenotypes for diabetes and obesity together.

Merck-internal vocabularies: TGI: Merck maintains a Target-Gene Information (TGI) system which includes a database of text-mined and SME-curated binary associations between genes and other biological entities (e.g., between "DGAT1" and "Adipoq"; "Insulin Resistance"; "fatty acid"; "Body mass"; …). From this database we extracted 13,863 binary associations (de-duplicated for case and directionality).

UMLS

We created a version of the UMLS Metathesaurus MRREL (relationship) file (2009AA release) with the terms mapped to the numerical concept identifiers, and from it extracted 205 relationships encoded by different UMLS source vocabularies for the 27 Merck D&O biomarkers and their GenBank synonyms/hyponyms (Table 2).

Sources, numbers, and examples (***concept1***) of benchmark D&O biomarkers/phenotypes extracted from UMLS (CUI: Concept Unique Identifier, RO: Related Other, RN: Related Narrow).

Gene ontology

We extracted the Gene Ontology (GO) primary relations of the 27 Merck D&O biomarkers and their GenBank synonyms/hyponyms using the GO Online SQL Environment [26]. A primary GO relation involves the GO annotations of the gene itself; for example, {"PRKAA1", *molecular_function*, "ATP binding"} or {"PRKAA1", *biological_process*, "fatty acid oxidation"}. Secondary relations were then computed by matching the primary GO terms to a downloaded version of GO. For example, since "PRKAA1" is annotated with "fatty acid oxidation" it would pick up a secondary relation to "fatty acid metabolic process" by virtue of the internal GO relation {"fatty acid oxidation", *is_a*, "fatty acid metabolic process"}. The result was 4,104 primary and 3,688 secondary GO reference D&O biomarkers/phenotypes.

OMIM

Disease-gene links in the Online Mendelian Inheritance in Man (OMIM) database were manually extracted for the 27 Merck D&O biomarkers and their GenBank synonyms/hyponyms, yielding 41 reference biomarkers/phenotypes, such as:

- D&O biomarker/hyponym: MC4R
- OMIM gene ID: 155541
- OMIM disease ID: 601665
- Disease name: OBESITY; LEANNESS, INCLUDED
- Disease-gene links: OB4, OB10Q, PPARGC1B, FTO, BMIQ8, GHRL, SDC3, …

Merck SME qualitative analysis

Of particular interest to Merck was the question "What biomarker/phenotype terms could be identified by the semantic engine that are in the Merck internal research documents and not publicly available in PubMed and ClinicalTrials.gov?" Creating this "unique to Merck" list was an exercise in cross referencing the three engine-produced lists for PubMed, ClinicalTrials.gov, and Merck internal research documents to uncover the terms in one list (Merck internal research documents) that are not in the other two lists (PubMed and ClinicalTrials.gov). The complete "unique to Merck" list was then culled of terms that were clearly not biomarkers/phenotypes and/or too general to be considered valuable medical terms.

InfoCodex output

The InfoCodex output was transformed into lists of D&O biomarker/phenotype candidates with their confidence level (CL) scores and other metadata. A total of 4,467 {*entity, biomarker/phenotype, diabetes/obesity*} candidate triples were found (1,361 and 1,743 biomarkers for diabetes and obesity, respectively, and 653 and 710 phenotypes for diabetes and obesity, respectively) ranging in CL from 3% to 70%, and distributed as shown in Figure 2. The highest scoring candidates discovered by InfoCodex text mining of the experimental PubMed collection are shown in Table 3.

Highest confidence level scoring biomarker/phenotype candidates discovered by InfoCodex text mining of the experimental PubMed collection. The identified candidate terms appear in column A, with their relationship to diabetes or obesity in columns B-C. The confidence level, in column D (the descending sort key), is normalized on a scale in which the maximum of 100% is the score of the manually curated reference biomarkers/phenotypes. In column E are the numbers

Source	*#rels*	*CUI-1*	*concept1*	*rel*	*relationship*	*CUI-2*	*concept2*
NCI	58	C0007595	FABP4 gene	RO	gene_plays_role_in_process	C1333527	Cell Growth
MSH	45	C0022621	FTO protein, mouse	RN	mapped_to	C2002654	Oxo-Acid-Lyases
OMIM	44	C0064317	KHK gene	RO	related_to	C1416630	Ketohexo-kinase
MTH	38	C0061352	GCGR gene	RO		C1415011	Glucagon Receptor
LNC	20	C0005767	MC4R gene mutation analysis:…	RO	has_system	C1715956	Blood

Table 2: UMLS benchmark sources, numbers, and examples.

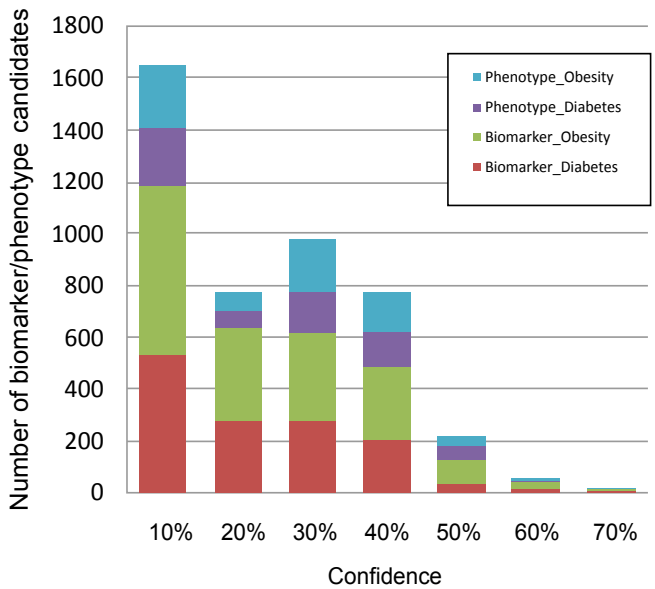

Figure 2: PubMed results confidence level distribution. Confidence level distribution of candidates discovered by InfoCodex text mining of the experimental PubMed collection.

of documents in which a given candidate term appears. Column F displays the PubMed IDs of the most relevant PubMed documents for purposes of manual SME review. Note that the same term can have multiple entries since it can have different relationships (biomarker for diabetes, phenotype for obesity, etc.).

Precision/recall

The fine conceptual/definitional difference between "biomarkers" and "phenotypes" was evident in the high degree of overlap in the two subsets produced by InfoCodex and I2E. Therefore we combined them for purposes of computing precision and recall. The results are shown in Table 4. The numbers tend to be low but there were some encouraging trends. InfoCodex precision/recall was higher for the more reliable manually parsed I2E output than for raw or auto-normalized I2E output, and could be made even higher by principled lumping of I2E terms (e.g., lumping *hyperglycemia, postprandial hyperglycemia, chronic hyperglycemia, hyperglycemia in women,* etc.). The high-end of the recall score ranges had good consistency for the most reliable benchmarks (I2E manual 33%, UMLS + GO + OMIM 35%, Thomson Reuters 36%).

Precision and recall of InfoCodex candidate biomarkers/phenotypes compared to various benchmarks. "(exact)" and "(preferred terms)" refer to sub-ranges according the 2x2 matching matrix described in the text under "Methods – Precision/recall". "MDOB" refers to the InfoCodex output subset containing references to the 27 Merck D&O biomarkers. "(unary)" means all InfoCodex candidate biomarkers/phenotypes were lumped together across obesity, diabetes, and MDOB, in contrast to the default binary criterion for matching.

The precision scores for individual biomarkers were highly variable, but some were impressive (I2E manual 52%, Thomson Reuters 49%, TGI 35%, ClinicalTrials.gov 59%) (not shown). For diabetes, there was a slight correlation between InfoCodex confidence level (CL) scores and precision against the I2E-manual benchmark (Figure 3). However, among the novel subset, there appeared to be a slight *inverse* correlation between quality and CL.

Novelty quality

Novelty is the "flip side" of precision; the "bad news" of low precision is accompanied by the "good news" of high novelty. But novel biomarker/phenotype candidates are useful only if they are of high quality (credible enough to justify follow-up research). Row 18 ("stimulant") in Table 3 and "antagonist" and "hypodermic" in Figure 3 would appear to be examples of low quality candidates. On the contrary, "insulin" (Row 2 in Table 3) and "proinsulin" (Row 3 in Table 3) are positive examples of proper candidates recognized as known biological complexes of diabetes.

Associative retrieval of known D&O biomarkers/phenotypes

In an effort to exemplify the associative recovery of a known phenotype of obesity, we used PubMed as a baseline to characterize the retrieval of a term InfoCodex specified as a phenotype. Melatonin receptor 1B (MTNR1B) is a candidate gene for type 2 diabetes acting through elevated fasting plasma glucose (FPG). As a phenotype of obesity, MTNR1B should not be considered novel, but it can be used

Row	Term (A)	Relationship (B)	Object (C)	Conf% (D)	#Docs (E)	PMIDs (F)
1	glycemic control	BiomarkerFor	Diabetes	70.3	1122	20110333, 20128112, 20149122,
2	Insulin	PhenoTypeOf	Diabetes	68.3	5000	19995096, 20017431, 20043582,
3	Proinsulin	BiomarkerFor	Diabetes	67.8	105	16108846, 9405904, 20139232,
4	TNF alpha inhibitor	PhenoTypeOf	Diabetes	67.1	245	9506740, 20025835, 20059414,
5	anhydroglucitol	BiomarkerFor	Diabetes	67.1	10	20424541, 20709052, 21357907,
6	linoleic acid	BiomarkerFor	Diabetes	67.1	61	20861175, 20846914, 15284064,
7	palmitic acid	BiomarkerFor	Diabetes	67.1	24	20861175, 20846914, 21437903,
8	pentosidine	BiomarkerFor	Diabetes	67.1	13	21447665, 21146883, 17898696,
9	uric acid	BiomarkerFor	Obesity	66.8	433	10726195, 19428063, 10904462,
10	proatrial natriuretic peptide	BiomarkerFor	Obesity	66.6	4	14769680, 18931036, 17351376,
11	ALT values	BiomarkerFor	Diabetes	66.3	2	20880180, 19010326
12	adrenomedullin	BiomarkerFor	Diabetes	64.3	7	21075100, 21408188, 20124980,
13	fructosamin	BiomarkerFor	Diabetes	64.2	59	20424541, 21054539, 18688079,
14	TNF alpha inhibitor	BiomarkerFor	Diabetes	62.1	245	9506740, 20025835, 20059414,
15	uric acid	BiomarkerFor	Diabetes	61.8	259	21431449, 20002472, 20413437,
16	monoclonal antibody	BiomarkerFor	Obesity	61.7	41	14715842, 21136440, 21042773,
17	Insulin level QTL	PhenoTypeOf	Obesity	61.2	1167	16614055, 19393079, 11093286,
18	stimulant	BiomarkerFor	Obesity	61.2	646	18407040, 18772043, 10082070,
19	IL-10	BiomarkerFor	Obesity	60.9	120	19798061, 19696761, 20190550,
20	central obesity	PhenoTypeOf	Diabetes	59.5	530	16099342, 17141913, 15942464,
21	lipid	BiomarkerFor	Obesity	59.5	4279	11596664, 12059988, 12379160,
22	urine albumin screening	BiomarkerFor	Diabetes	59.0	95	20886205, 19285607, 20299482,
23	tyrosine kinase inhibitor	BiomarkerFor	Obesity	58.8	83	18814184, 9538268, 15235125,
24	TNF alpha inhibitor	BiomarkerFor	Obesity	58.0	785	20143002, 20173393, 10227565,
25	fas	BiomarkerFor	Obesity	57.7	179	12716789, 17925465, 19301503,
26	leptin	PhenoTypeOf	Diabetes	57.6	870	11987032, 17372717, 18414479,
27	ALT values	BiomarkerFor	Obesity	57.4	8	16408483, 19010326, 17255837,
28	lipase	BiomarkerFor	Obesity	56.8	356	16752181, 17609260, 20512427,
29	insulin resistance	PhenoTypeOf	Obesity	55.8	5000	20452774, 20816595, 21114489,
30	chronic inflammation	PhenoTypeOf	Diabetes	55.7	154	15643475, 18673007, 18801863,

Table 3: PubMed results with highest confidence levels.

Benchmark	*Benchmark Corpus*	*InfoCodex Corpus*	*Precision*	*Recall*
I2E raw	PubMed	PubMed	(exact)	(exact)
			<1% obesity	5% obesity
			3-5% diabetes	9-11% diabetes
			3-7% MDOB	7% MDOB
I2E normalized	PubMed	PubMed	(exact)	(exact)
			3-7% MDOB	3-7% MDOB
I2E manual	PubMed	PubMed	1-5% obesity	9-33% obesity
			3-11% diabetes	9-31% diabetes
			3-26% MDOB	4-15% MDOB
UMLS + GO + OMIM	UMLS + GO + OMIM	PubMed	1-4%	3-22%
			1-8% (unary)	4-35% (unary)
Thomson Reuters	Thomson Reuters	PubMed	7-36% obesity	36% obesity
				18% DM2
			9-49% DM2	22% DM1
				25% DI
TGI	TGI	PubMed	0-5% obesity	(exact) 2.5%
			0-4% diabetes	
			1-14% MDOB	
I2E manual	PubMed	ClinicalTrials.gov	(preferred terms) 27-59%	(preferred terms) 3-7%
UMLS + GO + OMIM	UMLS + GO + OMIM	ClinicalTrials.gov	(preferred terms) 1-2%	(preferred terms) <1%
I2E manual	PubMed	Merck internal	(preferred terms) 8-14%	(preferred terms) 1-2%
UMLS + GO + OMIM	UMLS + GO + OMIM	Merck internal	(preferred terms) <1%	(preferred terms) <1%

Table 4: Precision and recall.

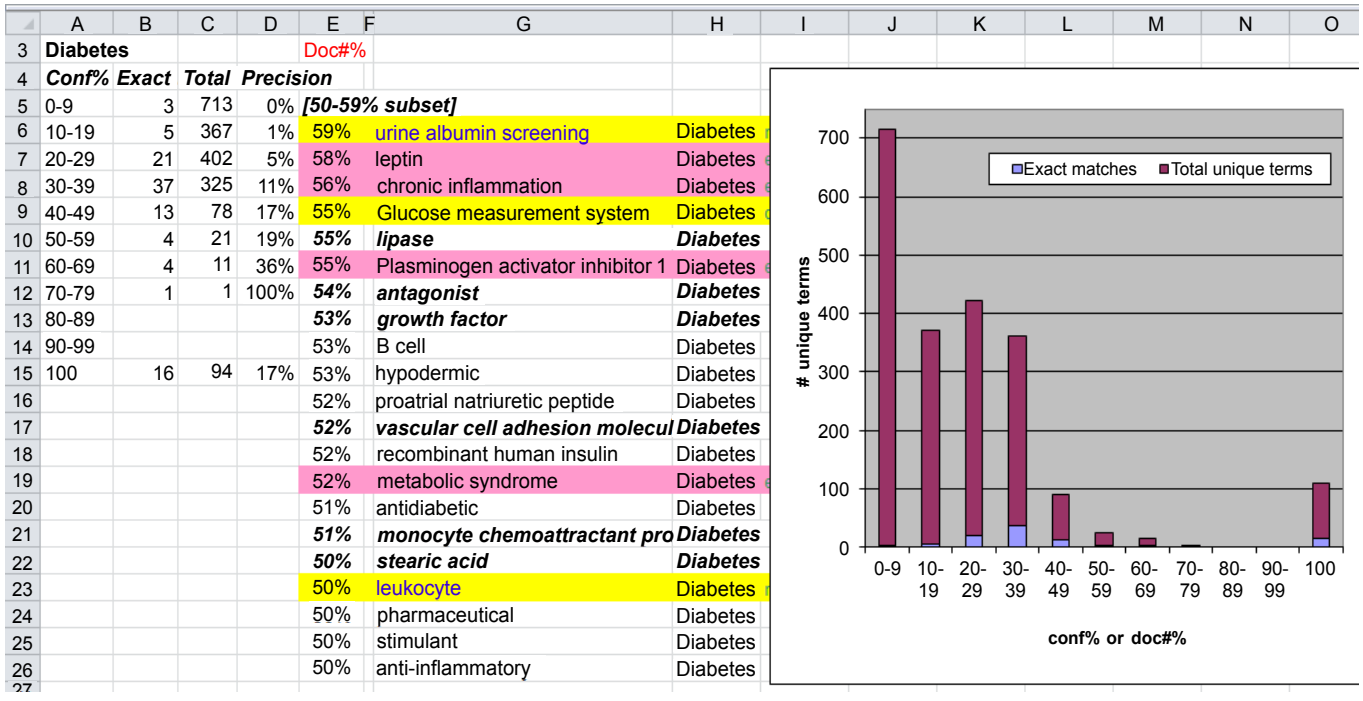

Figure 3: PubMed results confidence levels x I2E-manual precision.Correlation between InfoCodex confidence levels (Conf%; purple bars) and precision (light blue bars) against I2E-manual diabetes PubMed benchmark. Pink shading: exact match; yellow shading: partial match. Row 15 (100 Conf%) represents a member of the manually compiled reference set.

to substantiate the soundness of InfoCodex results extracted from PubMed and to illustrate the associative retrieval mechanism.

In PubMed, a search for "MTNR1B" AND "obesity" returned 9 documents, of which two (PMID: 20200315, 19088850) matched the PubMed abstracts selected by InfoCodex to substantiate its identification of MTNR1B as an obesity phenotype. When the criterion "phenotype" was added to the search, however, PubMed did not return any documents. A simple PubMed search would have thus failed to immediately identify MTNR1B as an obesity phenotype.

In PMID19088850, the word "phenotyping" is used to describe an action on a cohort of subjects, not a specification of MTNR1B as a phenotype. Later in the abstract the word "traits" is, however strongly indicating MTNR1B as a phenotype of obesity. The word "phenotype" is missing entirely in PMID 20200315. The InfoCodex semantic

engine could still correctly combine the MTNR1B-related information "increased prevalence of obesity" in PMID 20200315 with "traits" in PMID 19088850 to infer MTNR1B as a phenotype of obesity. A human read of these two abstracts would indeed immediately detect MTNR1B as a phenotype for obesity, identification the PubMed search engine failed to reveal, while the InfoCodex semantic engine was able to reconstruct it by integrating information distributed over the two documents even if the exact word "phenotype" never appears in relation to MTNR1B. Two abstract subsequently indexed by PubMed also fully confirm the identification of MTNR1B as a phenotype for obesity.

Thomson Reuter's relevance analysis

Thomson Reuters D&O SME analysts narrowed 2,369 (93%) novel obesity biomarker candidates down to 512 (20%) credible molecular biomarkers, of which 71 (3%) appeared to be initially confirmed by their presence on the Thomson Reuters Obesity Pathway Maps. For the finer relevance analysis, random samples of high- and low-confidence level InfoCodex/PubMed biomarker candidates were scored on the relevance scale from 0 to 10 as shown below (several thresholds of the scale below 10 reflect main types of erroneous associations between found biomarkers and diseases and how close they are in our opinion to relevant and unambiguous relationships):

- 10 – totally relevant and unambiguous relationship
- 8–9 – relevant, but can be associated with a related term – disease subtypes, disease symptom or consequence, etc.
- 6–7 – relevant, but correlation is rather remote. For example, some drugs may be causing elevation of blood pressure and should be administered with caution in diabetes patients (but drug is not for diabetes)
- 4–5 – associated in a specific context or found only one record
- 1–3 – low level of association
- 0 – no association, or term is so general it is not going to make sense

From the results in Table 5, it can be seen that only the obesity/molecular samples had respectable average relevance scores (6.9 high confidence, 6.2 low confidence). DM2/molecular and obesity/non-molecular terms averaged around 3 for both low and high confidence. DM2/non-molecular and both classes of DM1 exhibited an *inverse* confidence score effect, averaging around 1 for high and 3.4 for low. The main reason for low scores of non-molecular biomarkers with high confidence scores is the high percentage of terms that were considered to be too general and received score of 0; for example, "drug delivery", "first-in-class", "genotyping" and others.

Scale is described in main text.

UMLS mapping

A second approach to assessing the quality of the novel InfoCodex biomarker/phenotype candidates was mapping them to UMLS by co-sorting with the full 2009AA UMLS English lexicon extracted from the MRCONSO file.

The results are shown in Table 6. The highest percentage of exact matches was found for the novel InfoCodex biomarker/phenotype candidates from ClinicalTrials.gov (52%), followed by PubMed (39%), and lastly by Merck internal research documents. This order "makes sense" because new knowledge generally takes time to become canonical enough for controlled vocabularies. Clinical trials would be expected to be founded on the oldest, most well-developed knowledge, while Merck internal research concerns the newest and most tentative, with published literature being intermediate, consistent with our UMLS exact match results.

UMLS match type distribution of novel InfoCodex biomarker/phenotype candidates from the three corpora analyzed.

Merck SME qualitative results

10,953 novel biomarker/phenotype candidate terms were identified by InfoCodex from PubMed, ClinicalTrials.gov, and Merck internal research documents ("P3" in the figures). The summary for each data source and the overlap across data sources is summarized in Figure 4. Note that the overlap between ClinicalTrials.gov and Merck internal research documents "P3" is too small (i.e., count of 5) to be visible on this Figure 4.

Table 7 shows some examples of novel InfoCodex biomarker/phenotype candidates from Merck internal research documents that were clearly not biomarkers/phenotypes and/or too general to be considered valuable medical terms. The terms "wenqing" and "muise" are researcher names, and "shrna" stands for short hairpin RNA. Confidence levels reach the 50% + range in the example presented.

In general, high confidence levels and document counts characterize well-known biomarkers, as could have been expected. In addition to these, tens of interesting, plausible biomarkers/ phenotypes were found (not shown due to proprietary nature) by SME result assessment. These were concentrated in Merck internal research documents database (P3) but not in PubMed or ClinicalTrials.gov and are typically expressed with low CLs (<15%) and document counts (<7). While the low document count is fully understandable, the low confidence score of these potentially very interesting candidates is due to an erroneous inclusion of the document count in its definition.

Examples of uninteresting novel InfoCodex biomarker/phenotype candidates from Merck internal research documents.The terms "wenqing" and "muise" are researcher names, "shrna" stands for short hairpin RNA.

Discussion

One of the major high-level novelties of this experiment with

Biomarker type / Disease	*Average relevance scores for high confidence candidates*	*Average relevance scores for low confidence candidates*
Molecular Biomarkers		
Diabetes Type 1	1.6	3.2
Diabetes Type 2	3.6	3.7
Obesity	6.9	6.2
Non-molecular Biomarkers		
Diabetes Type 1	0.7	3.4
Diabetes Type 2	0.9	3.6
Obesity	2.6	2.8

Table 5: SME relevance analysis.

Corpus	*Exact*	*Left substring*	*Between 2*
Pubmed	789 (39%)	591 (29%)	632 (31%)
ClinicalTrials.gov	409 (52%)	225 (29%)	155 (20%)
Merck internal	24 (28%)	25 (29%)	38 (44%)

Table 6: UMLS match type distribution.

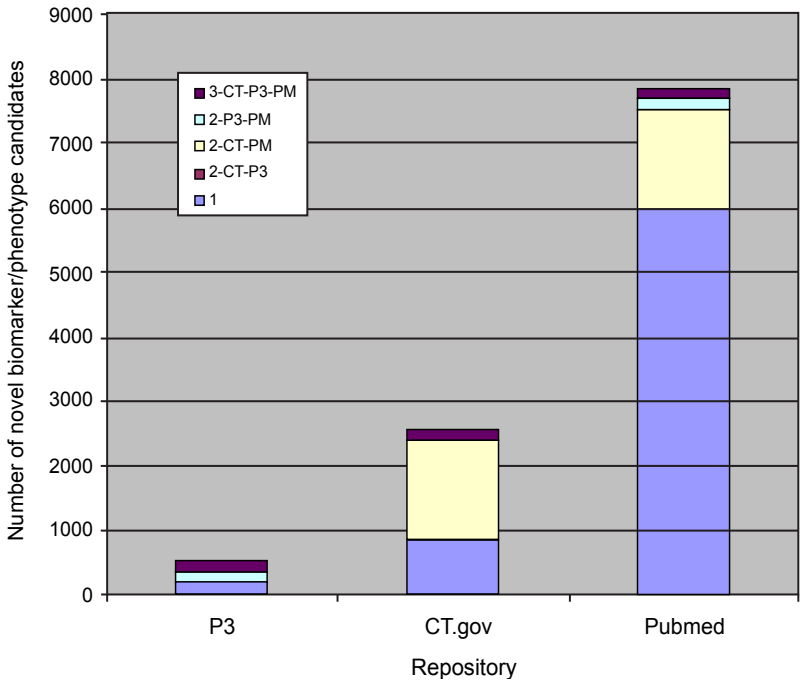

Figure 4: Novel candidates repository overlap. Overlap between novel InfoCodex biomarker/phenotype candidates from PubMed (PM), ClinicalTrials.gov (CT), and Merck internal research documents (P3). Lavender shading: found in one repository only; dark violet shading: found in all three; others: found in two (the overlap between CT and P3 is too small to be visible).

Term	*Relationship*	*Object*	*Target*	*Conf%*	*#Docs*
wenqing	Biomarker for	Obesity	Obesity	53.5	29
proteomic	Biomarker for	Obesity	Obesity	40.8	128
gene expression	Biomarker for	Obesity	Obesity	38.9	62
Mouse model	Biomarker for	Obesity	Obesity	19.8	17
muise	Biomarker for	Obesity	Obesity	17.5	20
athero-	Biomarker for	Obesity	Obesity	16.5	6
shrna	Biomarker for	Obesity	Obesity	9.6	4
inflammation	Biomarker for	Obesity	Obesity	8.2	4
TBD	Biomarker for	Obesity	Obesity	7.4	3
body weight	Phenotype of	Diabetes	MGAT2		1
cell line	Biomarker for	Diabetes	MGAT2		1

Table 7: UMLS benchmark sources, numbers, and examples.

respect to other recent studies [6] lies in the fact that the experiment was designed to test the power of autonomous self-organizing semantic engines. By design, the experiment was handled strictly as a "blind experiment" and no feedback from preliminary results was used to improve the machine-generated results.

Compared with recent studies [27-30] aimed at the extraction of drug–gene relations from the pharmacogenomic literature, this experiment introduces three novelties. First, while most related work is based on high-quality, manually curated knowledge bases such as PharmGKB [27] to train the recognition of connections between specific drugs and genes, our experiment's reference/training set (Step 1) was assembled in an *ad hoc* way by naïve (non-expert) PubMed search. Second, aside from the generic ontology in the ILD, no context-specific vocabularies (e.g., UMLS) were provided to inform the semantic engine. The meaning of unrecognized words had to be inferred by the InfoCodex engine based only on its universal internal linguistic database. Third, the text mining algorithms used here do not use rule-based approaches [29], or analyze co-occurrences sentence by sentence [27] or section by section [30], but rather they extract knowledge from entire documents and their relations with semantically related documents.

Natural language processing (NLP) approaches extract possible relations through analyzing documents sentence by sentence. Basically, such techniques can detect only those relations that have been written down by an author in some form or another, i.e., that are already known to some extent. Discoveries of really novel relationships require more than a sentence-by-sentence analysis. They are rather a result of the combination of small, seemingly unrelated and unnoticed facts dispersed over isolated publications. This is exactly what the InfoCodex approach intends to achieve, combining semantic technologies with statistical and neural analysis of whole document collections.

Among the discovered potential biomarkers/phenotypes there are some candidates of apparent high quality ("needles in the haystack"). Some of these have been tested, with encouraging results, for actual

novelty in a very preliminary way by internet searches (e.g., "xyz obesity" in Google or PubMed) where "xyz" is one of the candidates and "actual novelty" is defined as low hit rates, near or at zero, compared to known biomarkers (e.g., "adiponectin obesity"), with hit rates in the hundreds of thousands. More rigorous testing will require sizable effort and so we leave it for future follow-up studies.

Despite these successes, many results are not plausible or incompletely specified. This is not surprising for the following reasons:

- No prior knowledge on biomarkers/phenotypes was provided to the analysts who assembled the reference/training set (Step 1) and re-iteration was not allowed.
- Domain-specific knowledge (e.g., UMLS) was not added to the ILD to help the clustering or term extraction processes.
- Although it is certainly true that a large amount of human work was required to assess the quality of the generated results for potential novel biomarkers/ phenotypes in the proof-of-concept phase, the semantic analysis process for a discovery of novel biomarkers was largely automatic. No human expert feedback was allowed to influence the results. According to the blind nature of the experiment, the pure machine intelligence has been tested.

In view of these constraints, the capability of automatically identifying high quality candidates is very encouraging. The machine discovery process can deliver a list of potential biomarkers and can aid the biomarker discovery process by prioritizing them for follow-up research by confidence scores.

On the basis of the quality assessment by human SMEs, the quality of the machine discovery could substantially be improved by the following measures:

- Utilization of reliable SME-curated training sets of biomarkers/ phenotypes for the construction of the reference models (Step 1 above). In the present blind experiment the absence of any prior knowledge has led to a poor choice of some of the reference sets (e.g., generic terms such as "transforming growth factor" or "epistatic interaction" for biomarkers).
- Putting the focus of the novel biomarker discovery on proteins and genes as specified by the ILD ontology and giving other terms a lower weight.
- Extension of the ILD with additional proteins and genes taken from well-recognized biomedical dictionaries (e.g., UMLS), thus reducing the uncertainty in estimating the meaning of unknown terms and avoiding the use of incompletely specified terms.
- Use of named entity extraction rules to enhance the mapping of incomplete terms to complete, standardized biological terms.
- Improvement of the scoring method used in the estimation of the confidence level.

Conclusions

The reported approach of employing autonomous self-organizing semantic engines to aid biomarker discovery shows much promise and has potential to impact pharmaceutical research, for example to shorten time-to-market of novel drugs, or for early recognition of dead ends such as prohibitive side-effects through targeted extraction of relevant information.

The best approach to machine discovery must be considered as a semi-automatic, rather than a fully automatic, process since it cannot fully replace the competence of human researchers. The most promising approach is a hybrid process in which the automatically inferred discoveries are assessed by human experts.

In conclusion, we stress that what we presented here is a first step in an iterative process in which the machine discovery of biomarkers/ phenotypes and related pharmacogenomic entities is perfected to a level sufficient for human assessment of only the top tier of proposed novel entities. The final machine process we have in mind should not only lead to cost cutting with respect to traditional human research but it could become a valuable ingredient to tackle the sheer number of relevant documents available.

Competing Interests

The authors' corporate affiliations are given on the title page. Merck & Co., Inc., provided funding and computing resources for the work reported here.

Authors' Contributions

CAT co-designed the InfoCodex semantic engine and supported the development of specialized algorithms described under "Discovering biomarkers/ phenotypes with InfoCodex". DP was the Merck project leader and driving force. Both authors read and approved the final manuscript.

References

1. The changing role of chemistry in drug discovery: Thomson Reuters: International Year of Chemistry (IYC 2011) report.
2. Ranjan J (2005) Applications of data mining techniques in the pharmaceutical industry. Technol: J Theor Appl Inf 61–67.
3. Mattos N (2005) IBM study.
4. Lu Z (2011) PubMed and beyond: a survey of Web tools for searching biomedical literature. Database 2011: baq 036.
5. Hahn U, Cohen KB, Garten Y, Shah NH (2012) Mining the pharmacogenomics literature: a survey of the state of the art. Brief Bioin 13: 460–494.
6. Garten Y, Coulet A, Altman RB (2010) Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics 11: 1467–1489.
7. Trugenberger CA, Wälti C, Peregrim D, Sharp ME, Bureeva S (2013) Discovery of biomarkers and phenotypes by semantic technologies. BMC Bioinformatics 14: 51-68.
8. Biomarkers market discovery technologies (proteomics, genomics, imaging, bioinformatics), applications (drug discovery, personalized medicine, molecular diagnostics) & indications (cancer, cardiovascular & neural) - global trends & forecasts (2011–2020).
9. PubMed.
10. ClinicalTrials.gov.
11. UMLS.
12. Gene.
13. Gene Ontology.
14. OMIM.
15. Thomson Reuters.
16. Wälti P, Trugenberger CA, Cuypers F, Wälti C (2008) Sprach- und text-vorrichtung und entsprechendesverfahren, Patents EP1779271-B1 and US2007-0282598-A1/US2008-0215313-A1.
17. Cover TM, Thomas JA (2006) Elements of Information Theory. 2nd edition, Hoboken, John Wiley & Sons.
18. Kohonen T (2001) Self-Organizing Maps. 3rd edition, Berlin: Springer Verlag.
19. Fellbaum C (1998) WordNet: An Electronic Lexical Database. Cambridge MA, MIT Press.

20. Barry JM, Pollard JP, Wachspress EW (1989) A method of parallel iteration. J Comput Appl Math 28: 119–127.

21. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Statist 22: 79–87.

22. Shaw AP (2011) Semantic Tech & Business Conference: 26-27 September 2011, Trugenberger CA.

23. Späth H (1980) Cluster analysis algorithms for data reduction and classification of objects. Chichester: Ellis Horwood, Translated by Bull U.

24. Liu K, Hogan WR, Crowley RS (2011) Natural language processing methods and systems for biomedical ontology learning. J Biomed Inform 44: 163–179.

25. Linguamatics I2E.

26. GO Online SQL Environment.

27. Pakhomov S, Mcinnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, et al. (2012) Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. J Biomed Inform 45: 862–869.

28. Hakenberg J, Voronov D, Nguyen VH, Liang S, Anwar S, et al. (2012) A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. J Biomed Inform 45: 842–850.

29. Li J, Lu Z (2012) Systematic identification of pharmacogenomics information from clinical trials. J Biomed Inform 45: 870–878.

30. Xu R, Wang Q (2012) A knowledge-driven conditional approach to extract pharmacogenomics specific drug–gene relationships from free text. J Biomed Inform 45: 827–834.

Citation: Trugenberger CA, Peregrim D (2013) Discovery of Novel Biomarkers by Text Mining: A New Avenue for Drug Research? J Mol Biomark Diagn S3: 004. doi:10.4172/2155-9929.S3-004

This article was originally published in a special issue, **Biomarkers Discovery & Validation** handled by Editor(s). Dr. Krishhan VV, California State University, USA