Reconstruction and analysis of intercellular signaling networks

Dissertation zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.) der
Technischen Fakultät der Universität Bielefeld

vorgelegt von
Andre Skusa

Bielefeld im März 2006

Dipl. Inform. Andre Skusa
"Reconstruction and analysis of intercellular signaling networks"
Dissertation
Bielefeld 2006
Technische Fakultät der Universität Bielefeld
Gedruckt auf alterunbsbeständigem Papier (ISO 9706)

Zueignung

Ihr naht euch wieder, schwankende Gedanken,
Die früh sich einst dem trüben Blick gezeigt.
Weis' ich euch diesmal in die Schranken?
Sind Form und Inhalt recht bezeugt,
Wiewohl mein Herz dem wilden Ranken,
Ach! dem Mäandern zugeneigt?

Ihr drängt euch zu! nun gut, so mögt ihr walten,
Der Arbeit labyrinthisch irren Lauf
So rein und eben zu gestalten,
Dass man ihm folgen mag darauf.

(frei nach Goethe (1808))

Dank

Diese Arbeit hätte nicht geschrieben werden können, ohne die Unterstützung und tätige Hilfe vieler. An erster Stelle gebührt Dank meinen beiden Betreuern, Klaus Prank und Jens Stoye, die mir jederzeit mit Rat und Tat zur Seite standen. Ihre Offenheit der jeweils anderen Disziplin gegenüber ist eine wichtige Voraussetzung, um Methoden der Informatik auf biomedizinische Fragestellungen anwenden zu können.

Damit ein Informatiker ein bioinformatisches Thema bearbeiten kann, ist fachliche Unterstützung im biomedizinischen Bereich unumgänglich. In dieser Arbeit haben diesen Part Kai Lindemann und Daniel Klockenbrink, beide Mediziner von der Medizinischen Hochschule Hannover, übernommen. Ohne sie wäre vieles nicht möglich gewesen. Insbesondere die Auswertung der Ergebnisse und Hinweise auf Verbesserungen sind ihnen zu verdanken. Wo immer im Text von "biomedical experts" die Rede ist, sind Kai und Daniel gemeint oder waren beteiligt. Herzlichen Dank dafür.

Auf der informatischen Seite ergab eine glückliche Fügung, dass ich Jacob Köhler an der Uni Bielefeld just in dem Moment kennenlernte als ich überlegte, wie ein Vorgehen mit Text Mining aussehen könnte. Basierend auf Jacobs Ideen und auf Ergebnissen eines von ihm zuvor geleiteten Projektseminares konnten wir ONDEX gemeinsam mit Alexander Rüegg, ebenfalls aus Ralf Hofestädts AG "Bioinformatik und Medizinische Informatik", entwickeln. Alexander sei darüber hinaus gedankt für seine unermüdliche Hilfe in allen Datenbankfragen. Dank gilt auch den Studenten des Projektseminars "Informationsextraktion aus Biomedizinischen Texten", Jessica Butz, Marc Essmeier und Anja Friedrichsen, die erste Versuche mit Text Mining für Zell-Zell-Interaktionen unternahmen und mir wichtige Hinweise lieferten.

Carsten Drepper und Thomas Schmitt-John, beide zur Zeit dieser Arbeit an der Uni Bielefeld, verdanke ich eine erste Anwendung der extrahierten Zell-Zell-Relationen ebenso wie ihre Unterstützung bei der Auswertung der erzeugten Daten. Carsten und Thomas ermöglichten mir darüber hinaus tiefere Einblicke in biologische Zusammenhänge der interzellulären Kommunikation.

Zu guter Letzt, aber sicher nicht an letzter Stelle, möchte ich mich bei Dion Whitehead, Uni Münster, und Sven Rahmann, Uni Bielefeld, für alle Hilfe bei der Erstellung des vorliegenden Textes bedanken. Klaus Prank, Dirk Evers und der "NRW International Graduate School in Bioinformatics und Genome Research" sei gedankt für finanzielle und organisatorische Unterstützung, Ralf Hofestädt für die Zusammenarbeit mit seinen Mitarbeitern und die zur Verfgung gestellte Hardware. Der European Science Foundation danke ich für ein Kurzstipendium am Rothamsted Research Institute (RRI) in Harpenden, UK, und dem RRI danke ich für die freundliche Aufnahme bei mehreren Besuchen in der Gruppe von Jacob Köhler. Nicht zuletzt Jörg Böke und der syskoplan AG gebührt Dank für die Duldung nebenberuflicher Aktivitäten, damit diese Arbeit beendet werden konnte.

Abstract

Cells in the human body communicate over long distances via two systems, the humoral system and the neuronal system. The humoral system works via first messenger substances, such as hormones, cytokines and neurotransmitters, which are released into the blood. Biomedical knowledge on this kind of intercellular signaling is well established, but in contrast to signaling processes inside cells, not much of this knowledge exists in a form that is easily accessible for automated approaches, such as databases or ontologies. Most of what is known about extracellular signaling is stored in terms of natural language text in the scientific literature.

The present study aims at the reconstruction and analysis of cell-cell signaling pathways by applying automated approaches. Therefore, relevant data is extracted from molecular databases as well as from biomedical literature by applying concept based text mining. For this purpose, models and corresponding graph representations are developed to assemble intercellular signals from partial information since available data sources are scattered and incomplete. The resulting information is finally applied to generate hypotheses on cell-cell signaling in the context of neurodegenerative diseases.

More specifically, from the few molecular databases containing appropriate data, one database is tested in a preliminary study and reconstruction approaches accessing the specific structure of this database are developed. To reconstruct information from natural language text, ONDEX, a framework for ONtological text inDEXing and data integration has been developed in a collaborative work. ONDEX supports *concept based* approaches, i.e. databases and ontologies are integrated into a standardized graph-based framework, where biological entities as concepts are linked by relations (i.e., "is-a", "part-of" or "synonym"). A major part of this thesis is the development and the integration of concept based text indexing and concept based co-occurrence searches into ONDEX. On this basis, MEDLINE abstracts are mapped to concepts of a number of ontologies (e.g., Gene Ontology, MeSH terms and Cell Ontology) and mined for relevant parts of intercellular signaling. From these relations finally, cell-cell signaling hypotheses are assembled.

Whereas the networks resulting from the database reconstruction are not sufficient for reasonable analysis and further use, evaluations of the text mining results show that a significant number of known facts can be found by applying concept based co-occurrences searches. Finally, the text extraction results are reduced to a manageable amount of concept based co-occurrence hits and hypotheses for cell types involved in neurodegenerative diseases. In this case a number of known facts are reconstructed and suggestions for further improvements are made.

The text extraction results demonstrate the possibility to reconstruct relations between biological entities from text by applying a concept based framework and thus, how a large text set can be reduced to a number of hypotheses allowing manual examination.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Intercellular signaling

Cells are complex biological systems controlled by the interaction of molecules (Cooper, 2000). The state of a living cell is sustained by biomolecular networks on several levels, ranging from the regulation of gene expression to the control of energy consumption and production by metabolic networks (Barabási and Oltvai, 2004). In multicellular organisms, the actions inside single cells need to be coordinated and synchronized in order to shape the organism as a whole, coherent system (Alberts *et al.*, 2002). For example, immune system cells send signals to repel dangerous external intrusions. Another example is during ontogenesis: the cells need to "know" their special function and communicate for this purpose with each other.

Therefore, intercellular or extracellular signaling comes into play to connect cells. In contrast to intracellular signals inside a cell, an extracellular signal needs a messenger substance that is able to pass through the cell membrane, bridge a distance to another cell and to dock there, either at the cell surface or inside the target cell. With this mechanism, information can be transmitted between cells, and collective behavior can be initiated. Different cell types in an organism are able to send a variety of signals and can likewise react to signals in different ways. Thus, the cell types and their signaling capacities form a network where the cell types are the nodes connected by their signaling relations. Furthermore, the intercellular network of cell communication connects the intracellular networks and is hence an important mechanism to control cell function.

Networks are not only a metaphorical way to understand organisms as complex systems. A network perspective helps rather to organize knowledge on local interactions into a systemic view. Biology in general can be seen as a science of interconnected networks (Barabási and Oltvai, 2004). Systems biology, as a recently emerged branch of biological and life sciences, aims at assembling knowledge collected in specialized fields of molecular biology (Kitano, 2002; Hiesinger and Hassan, 2005). Additionally, network sciences attracted attention recently for the development of a theoretical base for a unifying description of systems as diverse as social networks, the internet or molecular interactions by using concepts from graph theory and statistical mechanics (Watts and Strogatz, 1998;

Barabási and Albert, 1999; Albert and Barabási, 2002; Newman, 2003). A combination of network sciences and the systematic assembly of already established partial biological knowledge will help to complete our understanding of the functionality of whole organisms.

In an integrative model of mammalian organisms, the intercellular signaling network is an important part. In order to gain such a system level view, the existing knowledge has to be collected and combined. The present thesis contributes to this goal by the development of methods for extracellular network reconstruction from databases and literature as well as the analysis and application of the reconstructed data.

Relevant data for reconstruction

Surprisingly, although cell signaling is a well-established field in biomedicine, and knowledge on cell signals has existed for more than 100 years, the amount of relevant data accessible for automated software approaches is very low. Structured information on complete cell-cell signals (i.e. comprising of sender and target cell types as well as the messenger substances connecting them) is not available directly, but rather as partial information that has to be combined into complete cell-cell signals. For this purpose, a cell signaling model and corresponding graph representations that reflect the form of the available data are developed in this thesis.

The first kind of data source applied to reconstruct intercellular signaling networks are molecular databases. Here structured information on components of intercellular signals is available. However, a major disadvantage of such databases that emerged after preliminary studies is the non-specific definition of the locations of messenger and receptor substances. In the few databases that specify molecular locations at all, there is usually no distinction between e.g., cell types or organs. Furthermore, often cell types of interest for specific problems are not contained in the databases.

A general problem in reconstructing cell-cell signals is that all reconstructed signals are unvalidated hypotheses due to their generation from partial information. Validating these hypotheses is complicated by the fact that many of them might hold true, but have simply not yet been investigated experimentally. Additionally, the number of resulting hypotheses due to combinatorial explosion of the available components of intercellular signals is very large, even for only few cell types.

Since all these problems apply probably for any currently available database, a text mining approach on abstracts of biomedical journal papers has been developed. Although the cell-cell signals have to be reconstructed similarly by combining partial data, the advantage of text mining is that the entities of interest can be specified in advance. Thus, search lists with cell types, messengers and receptors are applied. Additionally, text mining results in a set of potentially relevant texts that would be difficult to find by manual search queries.

Therefore ONDEX (Köhler *et al.*, 2004), a system for data integration, text mining, network extraction and visualization is developed in cooperation with Jacob Köhler (Rothamsted Research, Harpenden, UK) and Alexander Rüegg (Bioinformatics and Medical Informatics Department, Bielefeld University). ONDEX is a general purpose framework and not

restricted to the reconstruction of cell-cell networks, but several ideas designed in the context of the present thesis could be generalized and integrated in this system. In ONDEX, *concept based approaches* are proposed in order to enable data handling on a semantic level. For this purpose *ontologies* are used as background knowledge to identify concepts in the texts. Then the annotated texts can be queried for concepts rather than searching only at the string level. Hence, with the ONDEX framework it is for instance possible to detect texts that contain synonyms, abbreviations or different spellings of the searched terms.

In this context, ONDEX is used to index selected MEDLINE abstracts by a set of relevant concepts (i.e., cell types, messengers and receptors). Subsequently, a concept based co-occurrence search is applied to identify relations between these concepts. Using the resulting partial information on intercellular signals, cell-cell signaling hypotheses are finally generated. Furthermore, sequentially applied refinement steps in the co-occurrence searches serve as filter in order to reduce the amount of extracted information.

The reconstructed cell signaling network data is finally applied in the context of neurodegenerative diseases. Specifically, biologists of the group of Thomas Schmitt-John at Bielefeld University are conducting research on the Amyotrophic Lateral Sclerosis (ALS) disease and its respective model organism, the wobbler mouse (Schmitt-John *et al.*, 2005). The question regarding intercellular signaling is to identify communication relations between four particular cell types of interest. For this purpose, the text mining results could be used, whereas the available databases did not contain signaling information on the cell types considered.

Thesis overview

The thesis is structured as follows: Section 2 introduces the necessary background, i.e. the biological function of intercellular signaling, available molecular databases, text mining methods for network extraction as well as a brief review of existing approaches for reconstructing and analyzing intercellular networks. The background section concludes with a discussion of implications that follow from the presented state of research and should be considered for reconstructing cell-cell signaling networks. General intercellular signaling models and corresponding graph representations which are used in all applied data sources are defined in Section 3.

Sections 4 and 5 present the reconstruction approaches and results gained from the preliminary database study and from text mining in biomedical abstracts, respectively. Both sections describe the specific properties of the respective data source, the reconstruction approach following from these properties and its implementation. Exemplary results are discussed. In Section 6 the reconstruction results from both types of data sources, databases and text, are inspected to be applied for the search for signals between cell types relevant in neurodegenerative diseases. The thesis concludes with a discussion of the results and an outlook to future work (Section 7).

The appendix contains further information on existing network extraction tools, the implementation of ONDEX as well as several lists with terms used for and resulting from the presented network reconstruction approaches.

Chapter 2

Background

Contents

This chapter is intended to give an overview of the biological background of intercellular signaling (Section 2.1) and of the available electronic resources, such as molecular databases

(Section 2.2) as well as the biomedical literature and the possibilities to extract networks from text (Section 2.3). Existing approaches to reconstruct and analyze intercellular signaling networks are reviewed in Section 2.4. The chapter concludes with implications for this thesis resulting from the presented background (Section 2.5).

2.1 Intercellular signaling

According to the fossil record, sophisticated unicellular organisms resembling present-day bacteria were present on earth for about 2.5 billion years before the first multicellular organisms appeared (Alberts *et al.*, 2002). One reason why multicellularity was so slow to evolve may have been related to the difficulty of developing the elaborate cell communication mechanisms required for a multicellular organization. Cells must be able to communicate with one another in complex ways if they are to be able to govern their own behavior for the benefit of the organism as a whole.

These communication mechanisms depend heavily on extracellular signal molecules, which are produced by cells to signal to their neighbors or to cells further away. Each cell depends on elaborate systems of proteins that enable it to respond to a particular subset of signals in a cell-specific way. These proteins include cell-surface receptor proteins, which bind the signal molecule, plus a variety of intracellular signaling proteins that distribute the signal to appropriate parts of the cell. Using these mechanisms, intercellular communication controls a variety of important cellular processes (Figure 2.1, left side)

In this section the basic principles of cell signaling are explained. Therefore Section 2.1.1 defines the general types of cell signaling that are of interest here. The sections that follow present the different types of signals (Section 2.1.2), messenger substances (Section 2.1.3) and receptors (Section 2.1.4) constituting intercellular signaling. All explanations in this section only briefly describe the biological background necessary for the focus of the present thesis. Further information can be found in Alberts *et al.* (2002) and Cooper (2000), on which this introductory section is based[1].

2.1.1 General principles of cell signaling

The general mechanisms of cellular communication can be compared with the electronic transmission of information (as e.g. in telephone calls): The sender emits an electric impulse which is transported through a medium (wire) and received by a target where the message is decoded and, in some cases, causes responses. Translated to biological terms, in multicellular-organisms cell signaling comprises of a sender and a target cell as well as first messenger substances carrying the information. The messengers are finally decoded and transformed by a receptor molecule on or inside the target cell into second messengers. Cell signaling is therefore processed in two stages:

[1] All figures from Alberts *et al.* (2002) are reproduced by permission of Garland Science/Taylor & Francis LLC

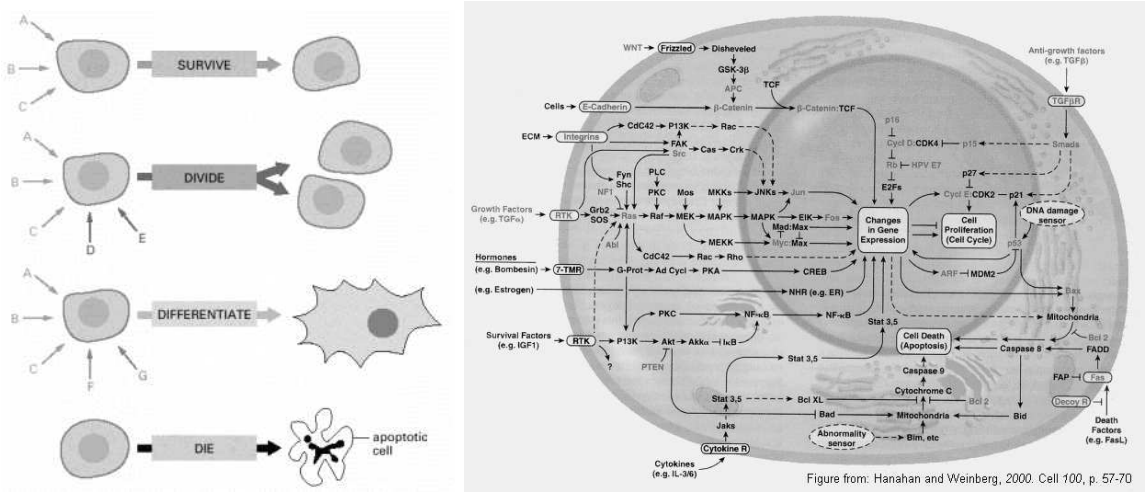

Figure 2.1: Intercellular and intracellular signals. **Left panel**: Possible effects of *inter*cellular communication. Each cell type displays a set of receptors that enables it to respond to a corresponding set of signal molecules produced by other cells. These signal molecules work in combinations to regulate the behavior of the cell. As shown here, an individual cell requires multiple signals to *survive* and additional signals to *divide* or *differentiate*. If deprived of appropriate survival signals, a cell will undergo a form of cell suicide known as *programmed cell death*, or apoptosis. **Right panel**: Growth signaling circuitry of the mammalian cell as example of *intra*cellular processes connected to extracellular signals. Genes highlighted in gray are known to be functionally altered in cancer cells (Sources: left figure copyright (©2002) from Alberts *et al.* (2002), right figure reprinted from Hanahan and Weinberg (2000) with permission from Elsevier).

1. *Signal transmission*: the *target cell* receives the information as *first messengers* (or *ligands*[2] released by a *source cell*). The first messengers bind to a specific *receptor* on the target cell (number (1) in Figure 2.2, left side).

2. *Signal transduction*: the ligand-receptor binding activates an intracellular signaling cascade of second messenger molecules (number (2) in Figure 2.2, left side). The transduction process "translates" the external signal so that cellular responses can take place.

The extracellular signaling molecules often act at very low concentrations, and the receptors that recognize them usually bind to them with high affinity. In most cases, these receptors are transmembrane proteins on the target cell surface. In other cases, the receptors are inside the target cell, and the signal molecule has to enter the cell to activate them: this requires that the signal molecules are sufficiently small and hydrophobic to diffuse across the plasma membrane. At the end of each intracellular signaling pathway are target proteins, which are altered when the pathway is active and change the behavior of the cell. Depending on the signal's effect, these target proteins can be for instance gene regulatory proteins, ion channels, components of a metabolic pathway or parts of the cytoskeleton (Figure 2.2, right side).

[2]both terms, ligand and first messenger, will be used interchangeably throughout this thesis

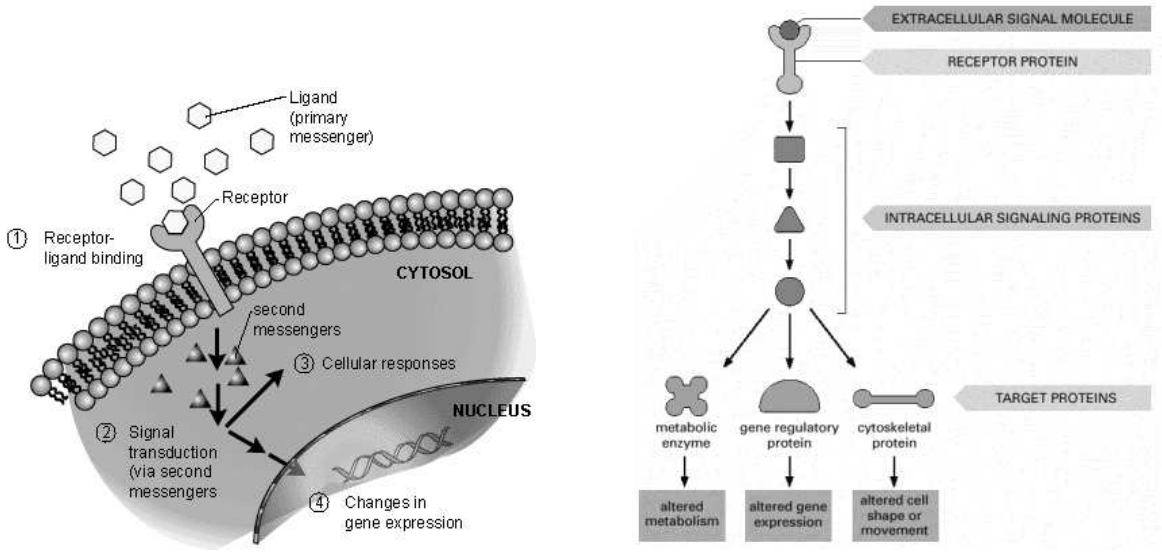

Figure 2.2: Principle mechanisms of cell signaling. **Left panel**: Ligands (or first messengers) are released by a source cell and bind to a receptor at a target cell (1). In the target cell the signal is transduced (2) and second messengers are released inside the cell, causing certain cellular responses (3) or altering the gene expression (4). **Right panel**: Schematic view of the same process. An extracellular signal molecule binds to a receptor protein and over several stages of an internal signaling cascade several different processes are activated or inhibited. (Sources: *BioTeach* (`www.bioteach.ubc.ca`) and copyright (©2002) from Alberts *et al.*, 2002).

The signaling cascade initialized by the ligand binding at the receptor is part of the complex *intra*cellular network that hierarchically combines the interactions of several classes of molecules inside a cell (from the level of genetic regulatory networks to protein-protein interactions and metabolic pathways, see Figure 2.1 (right side) as example).

2.1.2 Types of signals

Signal transmission between two cells differs mainly in respect to the cells' distance. The closest way of cell communication are cell *junctions* (see e.g. a *gap junction* in Figure 2.3, top left side). These cell-cell junctions can form between closely apposed plasma membranes and directly connect the cytoplasms of the joined cells via narrow water-filled channels. The channels allow the exchange of small intracellular signaling molecules (intracellular mediators), such as Ca^{2+} and cyclic AMP, but not of macromolecules, such as proteins or nucleic acids. Thus, cells connected by gap junctions can communicate with each other directly, without having to surmount the barrier presented by the intervening plasma membranes. Such junctions, however, will not be further considered here, since the present study focuses on cell signals based on ligand-receptor interactions.

In the closest ligand-receptor interaction based signaling type, the signal molecules remain bound to the surface of the signaling cell and influence only cells in contact to (Figure 2.3, left side, (A)). Such *contact-dependent signaling* is especially important dur-

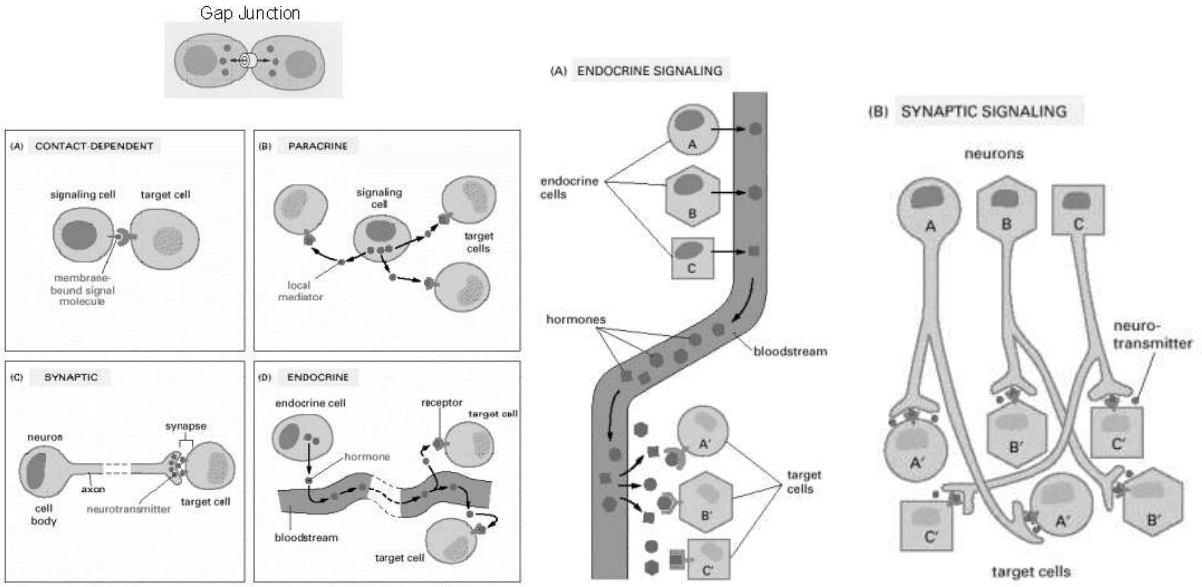

Figure 2.3: Types of cell signaling. **Left panel**: generally four different types of cell signals can be discriminated, varying from short range (A) to long distances (D). All these cell signals differ from cellular communication through direct *junctions*. For example, gap junctions (top) are specialized cell-cell junctions that can form between closely apposed plasma membranes and directly connect the cytoplasms of the joined cells via narrow water-filled channels. This way of cell communication is not based on ligand-receptor interactions and therefore not considered in the present study. **Middle and right panel**: schematic view of two selected signaling types for more than two connected cells. Whereas different endocrine cells (A) must use different hormones to communicate specifically with their target cells, different nerve cells (B) can use the same neurotransmitter and still communicate in a highly specific manner through the neuronal architecture (Source: copyright (©2002) from Alberts *et al.*, 2002).

ing development and in immune responses. In most cases, however, signal molecules are secreted and bridge a distance to the target cell. The secreted molecules may be carried far afield to act on distant targets, or they may act as local mediators, affecting only cells in the immediate environment of the signaling cell. This latter process is called *paracrine signaling* (Figure 2.3, left side, (B)). For paracrine signals to be delivered only to their proper target cells, the secreted molecules must not be allowed to diffuse too far; for this reason they are often rapidly taken up by neighboring target cells, destroyed by extracellular enzymes, or immobilized by the extracellular matrix.

For a large, complex multicellular organism, short-range signaling is not sufficient on its own to coordinate the behavior of its cells. In these organisms, sets of specialized cells have evolved with a specific role in communication between widely separate parts of the body. The most sophisticated of these are nerve cells, or neurons, which typically extend long processes (axons) that enable them to contact target cells far away through *synaptic signaling*. When activated by signals from the environment or from other nerve cells, a neuron sends electrical impulses (action potentials) rapidly along its axon; when such an impulse reaches the end of the axon, it causes the nerve terminals located there to secrete a chemical signal called a *neurotransmitter*. These signals are secreted at chemical

synapses, which are designed to ensure that the neurotransmitter is delivered specifically to the postsynaptic target cell (Figure 2.3, left side, (C)).

A second type of specialized signaling cell that controls the behavior of the organism as a whole is an endocrine cell. These cells secrete their signal molecules, called *hormones*, into the bloodstream, which carries the signal to target cells distributed widely throughout the body. These target cells have receptors for binding a specific hormone, which the cells "pull" from the extracellular fluid. This is called *endocrine signaling* (Figure 2.3, left side, (D)). In synaptic signaling, by contrast, specificity arises from the synaptic contacts between a nerve cell and the specific target cells it signals. Usually, only a target cell that is in synaptic communication with a nerve cell is exposed to the neurotransmitter released from the nerve terminal (although some neurotransmitters act in a paracrine mode, serving as local mediators that influence multiple target cells in the area).

In complex animals, endocrine cells and nerve cells work together to coordinate the diverse activities of the billions of cells. Whereas different endocrine cells must use different hormones to communicate specifically with their target cells (Figure 2.3, middle), different nerve cells can use the same neurotransmitter and still communicate in a highly specific manner (Figure 2.3, right side).

All of the forms of signaling discussed so far allow one cell to influence another. Often, the signaling cell and target are different cell types. Cells, however, can also send signals to other cells of the same type, as well as to themselves. In such *autocrine signaling*, a cell secretes signal molecules that can bind back to its own receptors. During development, for example, once a cell has been directed along a particular pathway of differentiation, it may begin to secrete autocrine signals to itself that reinforce this developmental decision.

2.1.3 Types of first messengers

According to the four general signaling types presented in the previous section, a possible classification scheme for first messenger substances is:

- *Contact-dependent signaling molecules*: an example for a signal molecule in *contact-dependent signaling* is *delta*, a transmembrane protein originating at prospective neurons and various other embryonic cell types. This messenger inhibits neighboring cells from becoming specialized in the same way as the signaling cell during development.

- *Local mediators*: in *paracrine signaling*, mainly growth factors act as messengers, such as e.g. the *epidermal growth factor* (EGF) or the *platelet-derived growth factor* (PDGF) that both stimulate many cell types to proliferate. A different example of a local mediator is *nitric oxide* (NO), a dissolved gas that is able to cross the plasma membrane of the target cell and directly binds to enzymes inside the cell in order to regulate smooth muscle contraction.

- *Neurotransmitters* are the first messengers in *synaptic signaling*. They diffuse across the synaptic cleft and bind to receptors on the target cell surface. Examples are

acetylcholine and *γ-aminobutyric acid* (GABA) which act excitatory and inhibitory respectively in the central nervous system.

- *Hormones* act as messengers in *endocrine signaling*. They can be divided in *peptide hormones* that bind to receptors at the cell surface and *steroid hormones* that cross the plasma membrane and bind to receptors inside the cell. Peptide hormones are for example *insulin* (stimulates glucose uptake), *glucagon* (stimulates glucose synthesis) and *growth hormones* (stimulation of several other substance and of the immune system). Exemplary steroid hormones are the sex steroids *testosterone*, *estrogen* and *progesterone* (induce and maintain secondary male/female sexual characteristics).

However, not any first messenger fits exactly into this scheme. Several signaling molecules exhibit the properties of more than one class, as e.g. *adrenaline* that increases blood pressure, heart rate and metabolism and acts as hormone as well as neurotransmitter.

A different way to classify extracellular signaling molecules is to divide them according to the two different fundamental types of receptors. The first and largest class of signals consists then of molecules that are too large or too hydrophilic to cross the plasma membrane of the target cell. The receptor proteins for these signal molecules therefore have to lie in the plasma membrane of the target cell and relay the message across the membrane (Figure 2.4, top left side). The second and smaller class consists of molecules that are sufficiently small and hydrophobic to diffuse across the plasma membrane. For these signal molecules the receptors lie in the interior of the target cell and are generally either gene regulatory proteins or enzymes (Figure 2.4, bottom left side).

Growth factors, neurotransmitters and peptide hormones belong to the first class of signal molecules that bind only to surface cell receptors. Growth factors are also known as *cytokines* which are mainly associated with hematopoietic (i.e., blood forming) cells and immune system cells (e.g., lymphocytes and tissue cells from spleen, thymus, and lymph nodes). Further members of this signaling molecule group are *chemokines* (a class of chemotactic cytokines) and *neuropeptides* (secreted by some neurons instead of the small-molecule neurotransmitters). The second group of signal molecules that bind to intracellular receptor is constituted by *steroid hormones* and the simple gas *nitric oxid* (NO).

2.1.4 Types of receptors

The main distinction that can be made for receptors is whether they are bound to the plasma membrane or reside inside the cell (Figure 2.4, left side). Inside these two groups further classifications can be shown:

Most *cell-surface* receptor proteins belong to one of three classes, defined by the transduction mechanism they use. *Ion-channel-linked receptors*, also known as transmitter-gated ion channels or ionotropic receptors, are involved in rapid synaptic signaling between electrically excitable cells (Figure 2.4, right side, (A)). This type of signaling is mediated by a small number of neurotransmitters that transiently open or close an ion channel formed

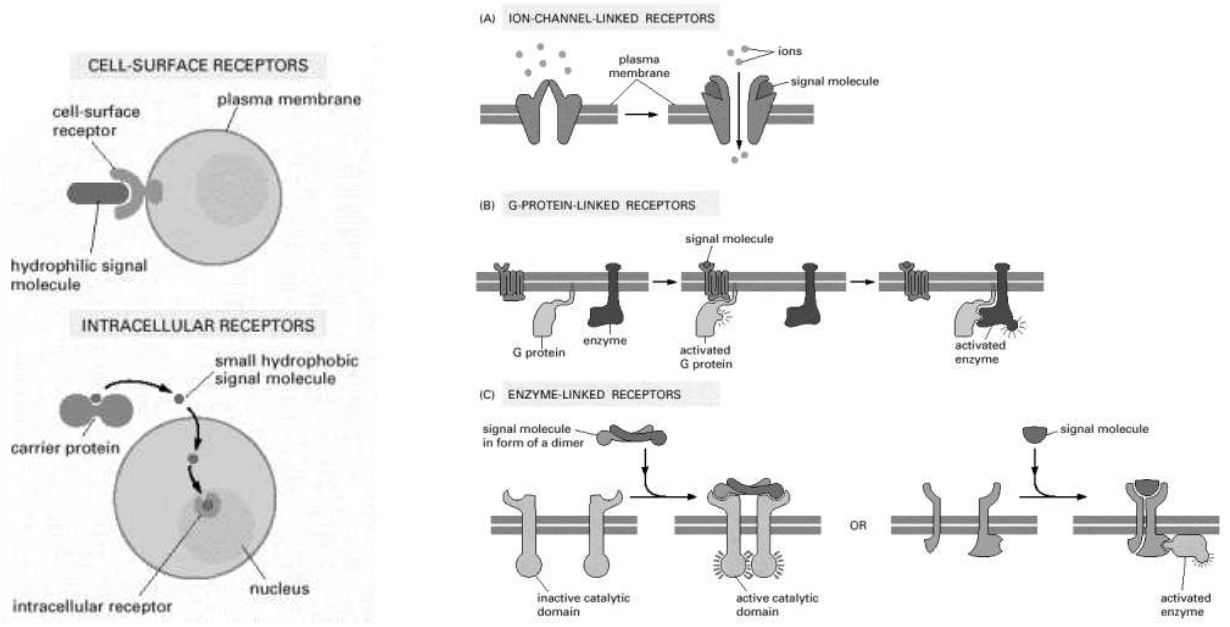

Figure 2.4: Receptor types in cell signaling. **Left panel**: the binding of extracellular signal molecules to either cell-surface receptors or intracellular receptors. Most signal molecules are hydrophilic and are therefore unable to cross the plasma membrane directly; instead, they bind to cell-surface receptors, which in turn generate one or more signals inside the target cell. Some small signal molecules, by contrast, diffuse across the plasma membrane and bind to receptors inside the target cell – either in the cytosol or in the nucleus (as shown here). **Right panel**: three classes of cell-surface receptors. (A) Ion-channel-linked receptors, (B) G-protein-linked receptors, and (C) enzyme-linked receptors. Although many enzyme-linked receptors have intrinsic enzyme activity, as shown on the left, many others rely on associated enzymes, as shown on the right (Source: copyright (©2002) from Alberts *et al.*, 2002).

by the protein to which they bind, briefly changing the ion permeability of the plasma membrane and thereby the excitability of the postsynaptic cell. The ion-channel-linked receptors belong to a large family of homologous, multipass transmembrane proteins.

G-protein-linked receptors act indirectly to regulate the activity of a separate plasma-membrane-bound target protein, which can be either an enzyme or an ion channel. The interaction between the receptor and this target protein is mediated by a third protein (G protein), called a trimeric GTP-binding protein (Figure 2.4, right side, (B)). The activation of the target protein can change the concentration of one or more intracellular mediators (if the target protein is an enzyme), or it can change the ion permeability of the plasma membrane (if the target protein is an ion channel). The intracellular mediators affected act in turn to alter the behavior of yet other signaling proteins in the cell. All of the G-protein-linked receptors belong to a large family of homologous, seven-pass transmembrane proteins.

Enzyme-linked receptors, when activated, either function directly as enzymes or are directly associated with enzymes that they activate (Figure 2.4, right side, (C)). They are formed by single-pass transmembrane proteins that have their ligand-binding site outside the cell and their catalytic or enzyme-binding site inside. Enzyme-linked receptors are

heterogeneous in structure compared with the other two classes. The great majority, however, are protein kinases, or are associated with protein kinases, and ligand binding to them causes the phosphorylation of specific sets of proteins in the target cell.

The *intracellular receptors* on the other hand all bind to specific DNA sequences adjacent to the genes the ligand regulates. Some receptors, such as those for cortisol, are located primarily in the cytosol and enter the nucleus after ligand binding; others, such as the thyroid and retinoid receptors, are bound to DNA in the nucleus even in the absence of ligand. The ligand binding also causes the receptor to bind to coactivator proteins that induce gene transcription. The transcriptional response usually takes place in successive steps: the direct activation of a small number of specific genes occurs within about 30 minutes and constitutes the primary response; the protein products of these genes in turn activate other genes to produce a delayed, secondary response; and so on. In this way, a simple hormonal trigger can cause a very complex change in the pattern of gene expression.

2.2 Molecular databases

The amount of biomedical data is increasing exponentially (Shatkay and Feldman, 2003). This is not only reflected by the large number of published journal articles in the respective research areas, but also by the accelerated growth of biomolecular databases. Furthermore, new databases for different purposes are frequently introduced.

In our case – the reconstruction of signaling interactions between cells – the situation is even more complex. After reviewing the contents of several available databases (Table 2.1), it was clear that no data source contains complete information on cell signaling, i.e. information of the form: cell type X sends messengers M to cell type Y. Instead, the most useful database content that can be found for our purpose is information on ligand-receptor interactions. From these interactions, cell-cell signals can be inferred by connecting the locations of the the ligand and the receptor molecule (for the biological background see the previous section). Therefore it is necessary that locations of the respective molecules are known, i.e. in which cell types ligands are produced or receptors are expressed.

Thus, the available databases are checked whether they contain interactions of the relevant molecules and their locations. The databases listed in Table 2.1 are selected exemplary to demonstrate the criteria used for choosing a data source.

Table 2.1 lists the databases according to their size (numbers of molecules and reactions contained, as far as current statistics are available). It should be noted that databases in some cases list *reactions* between the molecules (sometimes, as e.g. in KEGG, as chemical equations) and in some cases *interactions*. The exact chemical meaning could be different, but in our case the only information of interest is whether two substances are able to interact.

The databases differ not only in respect to their size, but also to their types: there are sequence databases (as Swiss-Prot or KEGG), containing mainly genetic information for a variety of organisms. Other databases focus on interactions (as the *Biomolecular INterac-*

Database	Mol.	Int.	Molecule locations?	Reference and URL
Swiss-Prot	195 589	–	Only as free text in the molecule comments	Gasteiger *et al.* (2001) `http://www.expasy.org/sprot`
Transpath	28 779	52 977	No locations given	Schacherer *et al.* (2001) `http://www.biobase.de`
HPRD	20 097	26 462	Expression sites given	Peri *et al.* (2003) `http://www.hprd.org`
DIP	18 827	55 393	No locations given	Xenarios *et al.* (2002) `http://dip.doe-mbi.ucla.edu`
KEGG (LIGAND)	13 042	6 442	No locations given	Kanehisa *et al.* (2004) `http://www.genome.jp/kegg`
CSNDB	3 512	1 382	Given for a subset of molecules	Igarashi and Kaminuma (1997) `http://geo.nihs.go.jp/csndb`
BIND	–	198 905	No locations given	Bader *et al.* (2001) `http://bind.ca`

Table 2.1: Statistics of molecular databases checked for information on intercellular signaling as of October 2005. The columns list the database name (1), the number of molecules (2) and interactions (3, in some cases referenced to as *reactions*), a comment whether cellular locations of the molecules are given (4) and the web address of the database (5). Further references are given in the text. The databases are listed in decreasing order of their molecule numbers.

tion Database (BIND) or the *Database of Interacting Proteins* (DIP)), signaling pathways (as Transpath or the Cell Signaling Network Data Base (CSNDB)) or on specific types of molecules (as the Human Protein Reference Database (HPRD)). All these databases contain molecules and interactions of interest regarding cell signaling, but they also exhibit one or several of the following problems:

- *Missing molecule locations*: the molecules are not assigned to cell types, tissues, organs or other anatomical locations, i.e. it is not known where they are synthesized.
- *Missing location types*: If locations are given, these locations are not further specified, i.e. it can not be determined automatically, whether the location is a cell type, a tissue or a different location type.
- *Missing molecule and interaction types*: Molecules and interactions are often not explicitly assigned to a type, as e.g. "ligand", "receptor" or "ligand-receptor binding". Thus, in such cases it can not be inferred only from the database which molecules and interactions are to be selected. Lists with molecules of interest are then required.

The most prominent problem in many databases is that molecule locations are not contained. The other two problems might be overcome by using additional data sources, as e.g. ontologies containing anatomical information or manually created lists with molecules of interest. Many databases contain also complete *pathways*, but except in the CSNDB these are intracellular pathways. Another restriction is that if databases contain only specific molecule types they might not cover all different first messengers of cell signaling.

In the following the databases shown in Table 2.1 will be briefly introduced (ordered as in the table). The focus of each description is the question whether and to which extent the respective database contains information of interest in respect to intercellular signaling. Further details of the databases are omitted here. Section 4 contains the reasons for the selection of the CSNDB and a more detailed description of this database, as well as the results of the applied reconstruction approaches. The present section closes with a brief review about ontologies that could be used as additional data sources to complement information missing in the molecular databases presented. Some of them are used later in the text mining approach (Section 5).

Swiss-Prot

Swiss-Prot (Gasteiger *et al.*, 2001) is a protein knowledge base established in 1986 and maintained collaboratively, since 1987, by the Swiss Institute of Bioinformatics and the European Molecular Biology Laboratory (EMBL). The database is part of the UniProt knowledge base, a central access point for curated protein information. Swiss-Prot is freely available and can be downloaded or accessed via a web interface. It is manually curated and aims at providing a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and a high level of integration with other databases (currently about 60).

The Swiss-Prot protein knowledge base consists of sequence entries, some intracellular signaling pathways, but no reactions of the protein encoded in the sequences. Thus, although Swiss-Prot is by far the largest database in our list (in respect to the number of contained molecules), it can not be applied to reconstruct cell signals.

Nevertheless, it was checked whether Swiss-Prot might be exploited to add information that is missing from other databases. However, Swiss-Prot is not of great use in this respect either: tissue locations are listed sometimes in the literature references or as free text in the comments of the molecule records. Also, all different kinds of locations (cell types etc.) are regarded as "tissue" and are not further specified. A further problem is the level of detail in Swiss-Prot. For example, if "insulin" is searched, the web engine lists 312 hits which include entries for different organisms and different forms as e.g. insulin precursors as well as receptors. Since the molecule types are not further specified (e.g. as "ligand" or as "receptor"), a parsing process would only work for pre-defined lists of entities.

For these reasons, Swiss-Prot is not further applied.

Transpath

Transpath (Schacherer *et al.*, 2001) has been developed and is supported as commercial database by the company Biobase, Wolfenbüttel, Germany. The database comprises of molecules participating in signal transduction and the reactions they undergo. Thus it spans the intracellular signaling network and together with the software PathwayBuilder, also developed and supported by Biobase, the overall intracellular network can be retrieved

and displayed.

Compared to Swiss-Prot, Transpath and the other selected databases are relatively small in respect to the number of molecules. But Transpath also contains reactions between its molecules as well as sequence information. Unfortunately, the only location information stored in Transpath are the intracellular locations of the molecules, not their tissues or cell types. Therefore, it could not be applied for database reconstruction. However, it can used as external data source and evaluation tool in the text mining approach (see Section 5.5).

HPRD

The Human Protein Reference Database (HPRD, Peri *et al.*, 2003) represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in the HPRD has been manually extracted from the literature by expert biologists. In only two years the database contents grew from 2 750 proteins and 10 534 interactions to currently 20 097 proteins and 26 462 interactions, the third largest database in our list (Table 2.1). Additionally, the HPRD provides an intuitive web engine to search and browse the contents and can be freely downloaded.

Although this database concentrates on intracellular signaling and proteins, it contains many of the molecules and interactions of interest in the context of extracellular signaling. Also, for each protein a list of expression sites is given, complemented by respective literature references. However, the exact type of the expression site (whether it is e.g. a cell type or a tissue) is not further specified.

In summary: the HPRD is a well-curated database that includes information on molecular locations in terms of expression sites. It might have been chosen instead of the CSNDB (see below) if it were available at that time, however, this database does not contain all necessary location information (especially the cell types needed for the application case in Section 6). Furthermore, even a database containing correct location information would not necessarily prevent the generation of very dense hypotheses networks, as demonstrated by the preliminary studies with the CSNDB.

DIP

The Database of Interacting Proteins (DIP, Xenarios *et al.*, 2002) is freely available and the fourth largest in our selection (Table 2.1), but more than 80% of the proteins are from non-mammalian organisms as *Drosophila*, *S. Cerevisiae*, *E.Coli* and *C. Elegans*. Only about 1000 reported proteins are from human, mouse and rat. Also, no locations of the molecules or interactions are stored. An example search for *insulin* returned the insulin precursor and the insulin receptor, but not the actual insulin hormone. Thus, this database is inappropriate for our purposes.

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) aims at enabling the computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and chemical information (Kanehisa *et al.*, 2004). It is freely available for searching and download. The sequence based part of the databases (GENES) contains more than one million entries from nearly 300 organisms and is thus even larger than the abovementioned Swiss-Prot. But in this context we refer to the LIGAND component of KEGG, containing about thirteen thousand molecules and six thousand reactions between them (Table 2.1). The reactions are denoted as chemical equations in text form as well as in graphical representations. Also figures for a large numbers of intracellular pathways are available. For these reasons, KEGG is one of the most frequently used sources for pathways, especially for enzyme related pathways.

However, locations of the molecules are not stored and hence, this database could not be applied in the context of extracellular signaling.

CSNDB

The Cell Signaling Networks Database (CSNDB) is designed as a data and knowledge base for signaling pathways of human cells. It compiles the information on biological molecules, sequences, structures, functions, and biological reactions which transfer the cellular signals (Igarashi and Kaminuma, 1997). The contents of the CSNDB are manually extracted from the scientific literature. It was freely available for searching, browsing and download, but is currently not online and seems to be not further supported.

The reasons to choose the CSNDB as test case of a extracellular signaling network reconstruction from a database are that molecular locations are defined and many relevant molecules are contained. Additionally, a number of extracellular signals are defined explicitly and molecules and interaction possess a type (as e.g., “hormone”, “cytokine” or “ligand-receptor binding”). Furthermore, the CSNDB mainly refers to the human organism.

A detailed description of the content and organization of the CSNDB as well as of the reconstruction results are the content of Section 4. A variety of cell signals could be extracted and verified, but the problems of databases in the context of extracellular signals (as listed above) remain.

BIND

The Biomolecular Interaction Network Database (BIND) is a collection of records documenting molecular interactions (Bader *et al.*, 2001). The contents of BIND include high-throughput data submissions and hand-curated information gathered from the scientific literature. BIND appears at the end of the database selection in Table 2.1 since it lists only the number of about 3 600 protein *complexes* in the database statistics, but not of the individual proteins contained.

BIND contains a considerable amount of molecules and interactions, detailed information are denoted in a new graphical notation called *ontoglyphs*, furthermore the database is freely available for searching and download but, however, no molecule locations are stored and thus, BIND can not be applied in our context.

Ontologies

In philosophy, *ontology* is the discipline considered with the study of *being* or *existence*. Therefore, an ontology defines basic categories that describe the nature and the organization of the world in an as much as possible objective way (in opposite to the subjective perspective of *epistemology*). In terms of computer science though, an ontology can be seen as "an explicit specification of a conceptualization" (Gruber, 1993), i.e. as a system for knowledge representation. Similar to expert systems they can be used to store facts about the world in a knowledge base and to define rules for inferring knowledge from the stored facts.

Ontologies can be briefly described as an extension of simple term collections or *controlled vocabularies* since ontologies additionally define relations between the entities (e.g. a *limb* can be characterized as *part-of* a *tree*). For detailed definitions of controlled vocabularies and ontologies see Section 5.1.1. Here we will only briefly mention some ontologies that could potentially be used to add information missing in the previously mentioned databases (as e.g. molecule types or specific information about the location of molecules). However, currently no single ontology provides sufficient information to completely reconstruct extracellular signals.

The *Medical Subject Headings* (MeSH) are part of the the Unified Medical Language System project (UMLS, Bodenreider, 2004) by National Library of Medicine (NLM) in the USA and used as controlled vocabulary for indexing articles in the MEDLINE database of journal abstracts. Each article contained in MEDLINE is manually assigned with a number of MeSH terms in order to characterize it and to improve database searches. Thus, the MeSH terminology aims at providing a consistent way to retrieve information that may use different terminology for the same concepts.

By using the hierarchical structure of the MeSH ontology, molecule names could be further characterized (as e.g. hormones or cytokines). In the context of text mining we used the MeSH terms in the opposite way to manually extract lists with the entities of interest (names of cell types, first messengers and receptors) that are to be searched in the texts (Section 5.2.1).

Ontologies reflecting the anatomical hierarchy in the human body could be considered in order to characterize missing molecular locations. A very comprehensive source in this context is the *Foundational Model of Anatomy* (FMA, Noy *et al.*, 2004, available at `sig.biostr.washington.edu/projects/fm/index.html`), a freely available domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy in a form that is understandable to humans and is also navigable by machine-based systems. However, the main assignments that could be made using the FMA are mappings

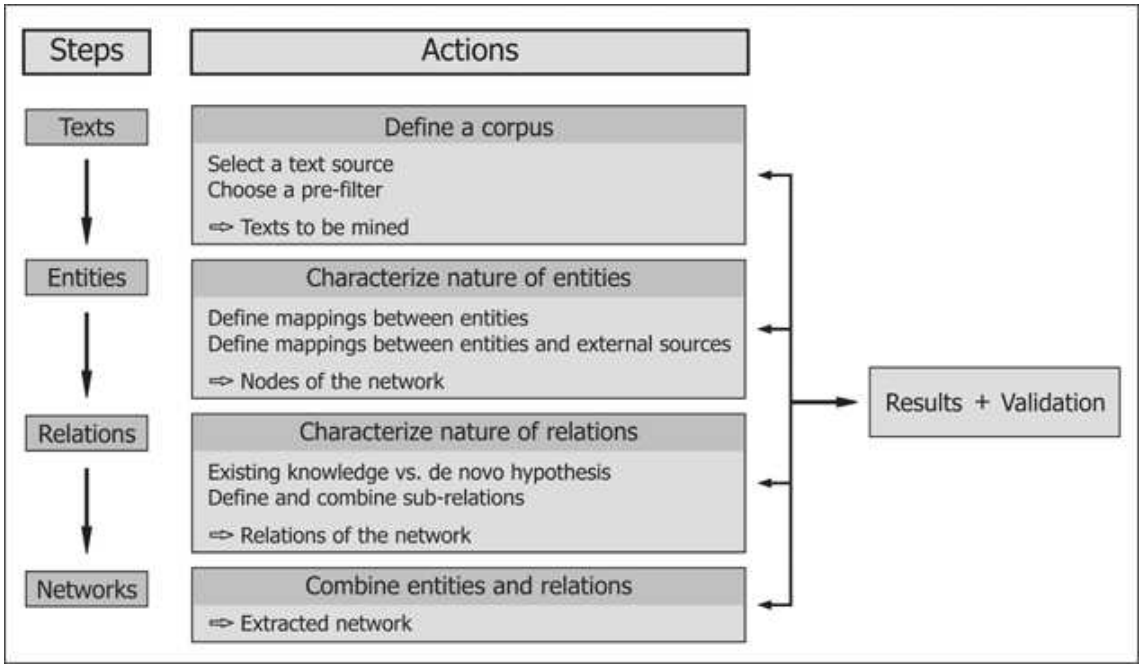

Figure 2.5: Schematic overview of approaches for network extraction from text.

to the major systems of the body, as e.g. the respiratory system or the cardiovascular system.

A further extensive anatomical ontology is the publicly available *eVOC ontology* (Kelso *et al.*, 2003, available at `www.evocontology.org`). eVOC is a controlled vocabulary that is based on molecular data and therefore unifies gene expression data by facilitating a link between the genome sequence and expression phenotype information. As a further source, *Cytomer* (Wingender, 2004) is a database describing a hierarchical classification structure for physiological systems, organs and cell types in order to enable the accurate description of regulatory events and expression patterns in organism, biological space and time. It is freely available from the Biobase company (`www.biobase.de`).

To conclude this section, the *Open Biomedical Ontologies* initiative should be mentioned (`obo.sourceforge.net`). Here a comprehensive and up-to-date list of controlled vocabularies and ontologies in the biomedical research is maintained.

2.3 Network extraction from text

Biology can be regarded as a science of networks: interactions between various biological entities (e.g. genes, proteins, metabolites) on different levels (e.g. gene regulation, cell signaling) can be represented as graphs and thus, analysis of such networks might shed new light on the function of biological systems (Barabási and Oltvai, 2004).

Such biological networks can be obtained from different sources. Especially the biomedical literature provides extensive and detailed information. Hence, the extraction of bio-

logical networks is an emerging text mining task, which requires the integration of a wide range of text mining techniques to support systems biological approaches in modeling, analysis and simulation of biological systems (Christopher *et al.*, 2004). Furthermore, network extraction is also important for other fields, such as database curation and annotation (Hofmann and Schomburg, 2005). Some databases such as Transpath (Schacherer *et al.*, 2001) are in fact networks, while others compile interactions between biological entities such as proteins, transcription factors or enzymes and metabolites, e.g. BIND (Bader *et al.*, 2001), DIP (Xenarios *et al.*, 2002), BRENDA (Schomburg *et al.*, 2002). Furthermore, extracted networks can be used to analyze and interpret experimental results, i.e. to support research and discovery (Werner, 2005). Another application is to exploit implicit information for generating new knowledge by combining extracted information into a set of hypotheses (Swanson, 1986; Srinivasan and Libbus, 2004; Wren *et al.*, 2004; Chen and Sharp, 2004; Eijk *et al.*, 2004).

The extraction of biological networks requires a combination of several different computational disciplines. Rather than presenting a comprehensive overview of each involved discipline or the whole relation mining field, this section aims at introducing key aspects and selecting examples that represent the different possible approaches.

Figure 2.5 introduces the main steps required for reconstructing biological networks from free text and serves also as guideline for this section: first the *texts* to be searched have to be chosen. Then *entities* (e.g. genes, proteins, metabolites) have to be identified and their (potential) relations are to be inferred from the selected texts. Finally, the entities and *relations* can be combined as nodes and edges into a *network*. The result produced in each step serves as input of the next step. Extracting structured information from unstructured natural language sources can not yet be expected to produce accurate results which can be used immediately and without further consideration. Therefore, the intermediate *results* of each step also deserve separate *validation* and their performance can be evaluated separately.

In the following, Section 2.3.1 introduces validation measures, the Sections 2.3.2 to 2.3.5 are organized along the steps presented in Figure 2.5. Section 2.3.6 closes this overview on network extraction from text with a summary. A brief survey about tools covering one or several steps of the workflow can be found in the appendix, Section A. This section is based on the publication "Extraction of biological interaction networks from scientific literature" (Skusa *et al.*, 2005). Further details can be found there.

2.3.1 Validation measures

For each step of the workflow (Figure 2.5), the performance is dependent on the previous steps. To quantify the performance of text mining results, three major metrics are normally used: recall, precision and effectiveness (Ding *et al.*, 2002). The *recall* is the fraction of correctly identified entities (texts, gene names, protein interactions, etc) in the set of *relevant* (i.e. true-positive) entities, whereas the *precision* is the proportion of extracted relevant entities to all entities retrieved. Precision and recall are sometimes also referred to as *specificity* and *sensitivity*. In simple words, the recall shows how much of the searched

information could be extracted and the precision reflects the *quality* of the method. From this it follows that in order to calculate the recall usually more information about the searched texts is needed in advance. On the other hand, in order to estimate the precision, one only has to validate a representative subset of the results obtained. For this reason, often the precision is reported without a recall. However, to provide a balanced estimate of the performance of a text mining approach, both values are combined in the effectiveness measure, which is the reciprocal of the mean of precision and recall.

2.3.2 Texts

The first decision to be made for the extraction of biological networks from scientific literature is the selection of the text sources. One drawback that can not be avoided is that even if relation mining would be 100% successful in retrieving all information from the respective literature, these networks would reflect mostly the current state of the literature, i.e. they might suffer from both the incompleteness and the biases of the current research efforts in molecular biology and genetics. In effect, networks extracted from scientific literature are not fully connected, and stronger connected subnetworks might stem from research activities concentrating on a couple of interesting genes or substances (Krauthammer *et al.*, 2002).

Although in principle any text source can be used for text mining, in practice abstract collections of scientific publications and full text journal publications are normally used. Abstract collections have the advantage of the relatively high information density. Further, they are often already manually annotated and categorized in a structured way that can be exploited for pre-filtering. Whereas MEDLINE (Bachrach and Charen, 1978) is the largest and most widely used bibliographic resource in the biological domain, other abstract collections and indexing services should also be considered, since MEDLINE does not necessarily provide the best domain coverage for a specific type of network to be extracted (Stone *et al.*, 1998). However, in most text mining approaches, MEDLINE is used, which is probably due to the fact that MEDLINE is freely available for noncommercial purposes.

Recently, an increasing number of text mining approaches also utilize full text journal publications (Friedman *et al.*, 2001; Huang *et al.*, 2004; Plake *et al.*, 2005), and the success of the open access model (Suber, 2002) will remove the financial hurdle for getting hold of a reasonable number of electronic full text publications. Yet dealing with full text publications is also more challenging on a technical level as one has to deal with a range of different formats (pdf, HTML) in which the publications are provided. The more demanding aspect is that the substructure is not always the same. However, since the typical sections of scientific publications (abstract, introduction, methods, results, discussion, figure captions, tables etc.) largely differ in their information density (Schuemie *et al.*, 2004), it is not surprising that those text mining applications applied on full texts perform best which take the substructure of the paper into account (Yeh *et al.*, 2003).

Once appropriate text sources are identified often the next step is to filter the text sources. In many cases, this is a simple need to reduce the amount of data into a man-

ageable subset: Mirroring and indexing all 15 million MEDLINE abstracts into a local database requires several days on a modern computer (Oliver *et al.*, 2004). The other reason for filtering is to improve the precision of the subsequent text mining steps by removing "obviously" irrelevant text sources. Often, simple methods (keywords, year of publication) are used for filtering. Yet there is the danger that such a simple approach may discard relevant texts. In order to define an organism specific filter for mice, a naive filter would be to only consider abstracts that contain the words "mouse" or "mice" or "mus musculus". However, such a filter will miss the 18000 MEDLINE abstracts with "murine" as the only word that indicates that they also refer to the same taxonomical entity. In other words, naive keyword filters may easily miss relevant information and thus already reduce the recall of the whole text mining process by filtering out relevant texts too early. For such reasons, advanced statistical and machine learning methods can be applied for pre-filtering (Blaschke *et al.*, 1999; Marcotte *et al.*, 2001).

In summary, the selection of the text sources and the definition of appropriate filters have a significant influence on subsequent steps – in the worst case, by selecting the wrong text sources or by applying the wrong filters even the best named entity recognition (NER, see "Entities", Section 2.3.3) and relation mining (see "Relations", Section 2.3.4) methods are deemed to fail.

2.3.3 Entities

Before relations can be searched for in texts, the entities of the relations have to be identified. Entities represent objects of the real world as e.g. proteins, genes, diseases etc. Usually these objects do not match simply to one name or symbol in natural language. Thus, many different words or symbols (as synonyms, abbreviations, acronyms or different spellings) have to be considered when a real world entity is searched in texts.

Named Entity Recognition (NER) is a longstanding NLP (Natural Language Processing) discipline on which a wide range of techniques exists. The different approaches and applications in bioinformatics are very well reviewed by Cohen and Hersh (2005) as well as by Krauthammer and Nenadic (2004). In the following, we will outline the basic ideas and principles.

According to Krauthammer and Nenadic (2004), NER consists of 3 steps: term recognition, term classification and term mapping, although term classification is not an important step for the purpose of network extraction from scientific literature.
For term recognition, the following approaches can be used:

- keywords: in the simplest case, lists of keywords are used to identify relevant entities.

- rules and regular expressions: for example entities such as fungal gene symbols, *Arabidopsis* gene symbols or enzyme numbers follow a standardized distinct syntax, that can reliably be extracted and identified by regular expressions (i.e. a string that describes or matches a set of strings, according to certain syntax rules). Yet, unfortunately not all taxonomical entities apply sensible genome nomenclature guidelines.

- dictionaries and ontologies: whereas dictionaries usually are used as simple term collections, ontologies also store typed relations between the terms, as e.g. "is a" or "part of" relations. Terms in ontologies are usually regarded as *concepts*. Entries in dictionaries and concepts of ontologies often contain several synonyms for the same entities. Dictionary based approaches can achieve a balanced precision and recall > 80% (Hanisch *et al.*, 2003; Ono *et al.*, 2001; Nenadic *et al.*, 2003). Thus, terminology and lexicon extraction from free text (Yu and Agichtein, 2003; Rindflesch *et al.*, 1999; Hatzivassiloglou and Weng, 2002) or from scientific databases (Chen *et al.*, 2005) are important disciplines in their own right. Another advantage of using dictionary based approaches is that the non-trivial task of term mapping (see below) becomes obsolete, and some dictionary based approaches can also be used for discriminating between different word senses (e.g. mouse as a pointing device versus an organism, see Hofmann and Schomburg, 2005; Ruch *et al.*, 2003). The ondex system, presented by Köhler *et al.* (2004) and in this thesis (Section 5.1) has been developed for this purpose as an integrated approach where ontologies and databases are mapped in order to perform concept based term identification and text indexing.

- machine learning: one of the most commonly used techniques is machine learning. Here, Support Vector Machines (Shi and Campagne, 2005; Kazama *et al.*, 2002) as well as Hidden Markov Models (Collier *et al.*, 2000; Shen *et al.*, 2003) are broadly and successfully applied.

Depending on the NER method used, equivalent entities are not always recognized as the same real world entity since, for most proteins and genes, several synonyms exist. Consequently, relationship mining methods that are developed on top of such NER methods would generate a good deal of redundancy. Such problems can be overcome by selecting an appropriate NER technique, or by subsequent computational or manual linkage of the equivalent entities (term mapping, see Wren *et al.*, 2005).

At the end of this step, the distinct entities (including in one entity all respective names, synonyms etc.) can be used as the *nodes* of the finally resulting network.

2.3.4 Relations

If the entities are defined and localized in the texts, relations between them can be inferred. Usually, the relations to extract are binary. They may or may not be directed or weighted with additional information. Furthermore, it is often required to determine the *type of the relation* (Smith *et al.*, 2005), e.g. whether they link proteins that *interact*, or whether they connect transcription factors that *regulate* genes. Most current efforts in relationship mining deal with protein-protein interactions: yet, also in these cases the different kinds of interactions (*activation*, *binding* etc.) need to be characterized.

Relation mining approaches range from applying simple statistical heuristics (e.g. by considering co-occurrences of search terms or estimating term frequency distributions) to syntactical and semantical sentence analysis (e.g. syntactical or semantical parsing)

using natural language processing (NLP) methods (Shatkay and Feldman, 2003). In *Rule-based approaches* a set of additional rules, that for example reflect prior experiences with the considered relation mining task, are added to improve the search (Yeh *et al.*, 2003). Furthermore, *machine learning methods* can be used to e.g. adapt patterns from text or to discriminate significant words (Plake *et al.*, 2005; Hatzivassiloglou and Weng, 2002).

Co-occurrence approaches

One of the most straightforward relation mining approaches is the *co-occurrence search*. The basic assumption here is that for describing a relation between two entities their names usually occur in the same text or part of the text. Thus, for co-occurring entities a relationship can be assumed.

Very basic approaches work with lists of keywords: For example a co-occurrence approach on the sentence level to search for nuclear receptors, their binding proteins and an interaction verb resulted in a precision of 22% when all extracted relations were examined manually (Albert *et al.*, 2003).

Another co-occurrence approach is applied in the PubGene database (Jenssen *et al.*, 2001) which contains gene-gene relations and was created by searching for pairs of gene names in MEDLINE abstracts. The extracted relations are weighted by the number of articles in which they were detected. Manual examination of two sets with each 500 randomly selected relations resulted in a precision of 60% for relations found in only 1 article and 71% for those found in 5 articles (recall not reported). Further evaluations were conducted by comparing the results with known gene-gene interactions from databases (DIP (Xenarios *et al.*, 2002), OMIM (Hamosh *et al.*, 2005)). Between 45% and 51% of the interactions in the database were also found by PubGene.

The performance of co-occurrence searches also depends on the part of the text in which co-occurrences are considered. Ding *et al.* (2002) compared recall, precision and effectiveness in single phrases, sentences or the whole abstracts. Interestingly, some *relation types* can best be extracted at the sentence level, whereas others perform better when whole abstracts are considered. Therefore, as a further enhancement, co-occurrence searches can be combined with a set of simple rules that determine the context size and order of the co-occurrence. For example, to extract protein-protein interactions (Blaschke *et al.*, 1999) in *Drosophila* the texts were divided into fragments (i.e. sentences or part of sentences). Then only co-occurrences of protein names and an interaction verb (all taken from predefined lists) possessing the form "protein A - verb - protein B" are extracted from these fragments.

Natural language processing approaches

Whereas in co-occurrence approaches only simple rules or patterns are applied to a small set of two or three extracted entities and additional words, *natural language processing* (NLP) techniques parse and analyze the sentences in greater detail (Manning and Schütze, 1999).

Shallow parsers (sometimes referred to as *partial parsers*) are used to identify the syntactic information that is assumed to be the most important. Here, mainly part-of-speech (POS) taggers are used for tagging each word in a sentence with its most likely grammatical function (e.g., noun, verb etc., see Manning and Schütze, 1999). This can then be used to infer the relations described (Leroy *et al.*, 2003; Sekimizu *et al.*, 1998). *Deep parsers* try to reconstruct the complete sentence structure as a tree structure (Daraselia *et al.*, 2004; McDonald *et al.*, 2004) and apply a grammar, such as e.g. the combinatory categorial grammar (CCG, see Park *et al.*, 2001), which first localizes target verbs to scan afterwards the neighborhood for the entities of the relations. Generally, full sentence parsers can be distinguished into such reconstructing the syntax or the semantics of a sentence, or a mixture of both. A review by McDonald *et al.* (2004) introduces both approaches and mixtures of them and gives an overview on applications in the biomedical text mining field and the resulting performances, advantages and drawbacks: Whilst syntax based approaches need no further domain specific information, they can easily be applied in different domains, but suffer from a lower precision than semantic parsers. For biological relation mining with one exception (Leroy *et al.* (2003) report 90%) no higher precision rates than 83% are reported. The only reported recall was about 47% (Yakushiji *et al.*, 2001). Contrarily, semantic grammars apply domain specific resources and thus result in an increased precision (up to 91% and 96%), but are often evaluated in a smaller sample of documents. Consequently, balanced or hybrid approaches have been developed, which try to exploit the benefits of both syntactic and semantic full parsing. The precision of such hybrid systems is high (e.g., 89% (McDonald *et al.*, 2004) or 91% (Daraselia *et al.*, 2004)), but the recall is still relatively low (35% (McDonald *et al.*, 2004) and 21% (Daraselia *et al.*, 2004) respectively).

Comparing NLP approaches with simple co-occurrence assumptions shows that NLP results in some cases in a higher precision, as one could expect from intensive grammar analyzes, but at the cost of speed and recall. On the other hand, NLP methods produce knowledge that can be exploited in steps which have to be performed separately when using co-occurrence searches. The POS tagging information can be, for example, used in the named entity recognition and the direction or the type of the relation can be easier inferred using the exact structure of the sentence.

Different relation mining strategies were compared in the "KDD Challenge Cup" (Yeh *et al.*, 2003). Despite the differences in their approaches, all winning teams have in common that they take the order of words into account rather than considering a text simply as a "bag of words". The fact that the winning team applied a purely rule based approach, and that the other top performing approaches also used a rule based component in their systems, indicates that machine learning approaches cannot yet compete with rules developed by experts.

Hypotheses generation

Relation mining as described so far can be characterized as reconstructing *established knowledge*, whereas other approaches try to generating *de novo* hypotheses by combining

extracted relations. Wren *et al.* (2004) and Srinivasan and Libbus (2004) both extend and improve the open discovery approach originally proposed by Swanson (1986). The basic assumption is that pairs of terms found in different texts and sharing the same "intermediate" terms can be linked.

An important improvement is to establish a robust and meaningful score for the extracted potential relations. Combining even only a few co-occurrence pairs usually results in a high number of possible implicit links. Wren *et al.* (2004) propose to use fuzzy logic methods and compare extracted networks with random networks. Srinivasan and Libbus (2004) use combined weights that rank the importance of each identified term (similar to the abovementioned scoring proposed by Stephens *et al.* (2001). In both papers hypotheses could be found that have not been reported in a single paper before and which led to new directions for experimental validation. Eijk *et al.* (2004) introduce the associative concept space (ACS) as metric for weighting the distance between pairs of terms according to the length of the chain of intermediate terms which connect them. Using this method, clusters of functionally related genes could be identified (Jelier *et al.*, 2005). In Chilibot (Chen and Sharp, 2004, see also the appendix, Section A) the whole extracted network is used to generate a network with hypothetical new interactions. Though, experimental validation is in most cases still the only way to prove the hypothesis.

As a result of relation mining, *links* of the network to be created can be gained. They might directly consist of a relation between two entities or consist of two or more combined relations.

2.3.5 Networks

Finally, the nodes and links created in the steps "Entities" and "Relations" can be integrated into a network. Yet such networks are incomplete and may contain incorrect entities and relations. As already discussed, in each of the different steps a range of methods can be applied that vary significantly in their precision and recall. Therefore, currently only very few approaches are published where networks extracted from texts are used for analysis and further investigations.

One possibility to deal with the uncertainty in the resulting networks is to apply a score that represents the quality of the extracted relations. Such a score can be used as an edge weight to visualize the likelihood of the correctness of relations. New discovered relations could be drawn in a different way (Blaschke *et al.*, 1999) and thus the network visualization can be used for manual comparison with existing knowledge by experts (Friedman *et al.*, 2001; Jenssen *et al.*, 2001; Yao *et al.*, 2004; Rzhetsky *et al.*, 2004).

In principle, extracted networks can be used for answering specific biologic questions or to provide deeper insights into the general structure of biochemical network topologies. In some cases the resulting network topologies have been investigated (Chen and Sharp, 2004; Blaschke and Valencia, 2001). Some topological characteristics of the network can be attributed to the bias of scientific literature (trendy topics and terms resulting in waves of publications on related genes, proteins etc., see Krauthammer *et al.*, 2002). But so

far, topological properties of hypothetical networks were mainly used for validating and analyzing the correctness of the extracted networks.

Rather than analyzing the topological properties, the extracted networks can also be used in context with experimental data in order to validate the extracted network as well as to evaluate the experiments. For example Jenssen *et al.* (2001) could show that their extracted co-occurrence gene networks reflect biologically meaningful relationships from three large-scale experiments. The resulting PubGene database and tool allows analysis of gene expression data in the context of extracted networks (see also the appendix, Section A). Karopka *et al.* (2004) apply their extraction approach on lists of gene names from experiments to compare extracted with experimentally determined relations. Albert *et al.* (2003) searched for protein interactions of nuclear receptors and compared these text mining results with data from yeast two-hybrid screens. Here they found similarities of the nuclear receptors regarding their connectivities. Also properties of some specific proteins were investigated and could be experimentally validated. Another example for the use of extracted networks is the curation of specific pathways, e.g. the Wnt pathway (Santos *et al.*, 2005).

2.3.6 Summary

Which presented extraction method performs best obviously depends highly on the specific types of networks to be extracted, and on the typical structure of a publication that contains a relation. For example, protein-protein interactions are often dealt with at the sentence level and achieve a good precision (up to 95%), but low recall in those few cases where the recall is also reported (Huang *et al.*, 2004; Daraselia *et al.*, 2004; Donaldson *et al.*, 2003). The type of networks to be extracted might also determine whether it is sufficient for the actual relation mining to use simple heuristics (as e.g. approaches based on co-occurrences of search terms in the same context) or whether there is a potential benefit in using advanced methods (such as e.g. syntactic or semantic parsing of sentences).

Although several systems exist that can be used for certain types of networks (mainly gene-gene and protein-protein interactions), a coherent "all-in-one" solution for extracting biological networks from text does not exist, nor is it appropriate to address the different types of problems in the same way.

An overview table (Figure 7.1) and more details about tools for network extraction and text mining that are applicable for all or individual steps of the workflow can be found in the appendix (Section A). Unfortunately, for several reasons none of these tools seem to be appropriate to use for the extraction of intercellular signals from text. In many cases the tools either are not available or require an additional database installation. Often the tools are specialized for the domain for which they have been developed. In other cases the tools are only available as web applications or are commercial. Hence, to develop ONDEX as a suite for concept based data integration, network extraction and visualization seems to be worth the effort. ONDEX and its application will be presented in Section 5.

2.4 Approaches to reconstruct and analyze intercellular signaling networks

The objective of this thesis is to reconstruct and analyze intercellular signaling networks. Most research conducted so far on biological networks concentrated mainly on *intra*cellular networks, such as genetic regulatory networks (Bower and Bolouri, 2000), metabolic networks (Ma and Zeng, 2003), protein-protein interaction networks (Schwikowski *et al.*, 2000) and signal transduction networks (Steffen *et al.*, 2002). Contrary, *inter*cellular signaling networks are in the focus of only a small number of research projects. Therefore we will introduce in the following such approaches.

2.4.1 Bioinformatics and cellular signaling

The general requirements to reconstruct and analyze cellular signaling networks are reviewed by Papin and Subramaniam (2004) and Papin *et al.* (2005). Here, cellular signaling is understood as integral combination of intra- and extracellular signals and thus, of events that happen at diverse spatio-temporal scales. Modeling cellular networks ranges from biochemical equations representing quick intracellular responses ($< 10^{-1}$ seconds, such as e.g. protein modifications and changes in Ca^{2+} concentrations) to slow responses (from minutes to hours) over large distances, such as in endocrine signals. The networks can be modeled in varying degrees of detail to understand their complexity and to make quantitative predictions. But a whole-network reconstruction at the most detailed level of differential equations or stochastic simulations is certainly out of reach. Thus, different kinds of modeling approaches have to be combined in order to gain a systemic view on cellular signals between cells.

In addition to the details on intracellular networks, on which the review by Papin *et al.* (2005) mainly concentrates, combinatorial calculations are presented that elucidate the complexity that a complete intra-to-extracellular signaling network would exhibit. Even if all intracellular elements that can influence the signaling processes are not considered, the variety of the possible ligand-receptor interactions is large. For example, 367 variants of the G-protein-coupled receptor (GPCR) could be identified in the human genome and the expression profiles of 100 GPCRs in the mouse genome also indicate that most receptors are expressed in various tissues. Hence, many different receptors probably exist concurrently in the same cell or tissue. If one assumes that a mere 1% of the estimated 1 543 different receptors in the human genome (i.e. 15 receptors) can be independently expressed in any give cell type, then a cell could potentially respond to $2^{15} = 32\,768$ different ligand combinations (for two independent ligand states: bound and unbound). This gives a good illustration of the general complexity of cellular signaling.

To finally reach the goal of an integrated model of the human cell signaling, efforts are needed that go beyond single research projects. For this purpose, several research initiatives have been formed to build the base for a human whole signaling network as well as for the integration of data at several physiological levels. There is for example the *Alliance*

for Cellular Signaling (Gilman *et al.*, 2002, available at `www.signaling-gateway.org`), which is a large-scale collaboration designed to answer global questions about signaling networks. But although this initiative addresses cell signaling in general, the effort is mainly restricted to intercellular signaling and there on pathways of two cells, *B lymphocytes* and *cardiac myocytes*. Further projects in this context are the *Database of Quantitative Cellular Signaling* (DOQCS, a repository of models of signaling pathways at the level of chemical reactions, Sivakumaran *et al.*, 2003, available at `doqcs.ncbs.res.in`) and the portal of the Science journal, the *Signal Transduction Knowledge Environment* (STKE, available at `stke.sciencemag.org`) . The STKE includes *Connection Maps*, the database of cell signaling. The integration of different and separately stored pathways at the intracellular level has been shown by Hsing *et al.* (2004). For this purpose they used *semantic networks* which are similar to *ontologies* (Section 5.1).

As systems biology emerged as a discipline with the goal to integrate existing knowledge from different levels of molecular biology (Kitano, 2002), an integrative modeling of all physiological levels in the human organism is achieved in two ambitious projects, the *Physiome Project* located at the University of Washington, USA (Bassingthwaighte, 1995, available at `www.physiome.org`), and the *IUPS Physiome Project* at the University of Auckland, New Zealand (Hunter *et al.*, 2005, available at `wwww.bioeng.auckland.ac.nz/physiome/physiome_project.php`). The physiome projects are worldwide efforts to define and describe the physiome quantitatively through the development of databases and models which will facilitate the understanding of the integrative function of cells, organs, and organisms. The aim is to develop integrative models at all levels of biological organization, from genes to the whole organism via gene regulatory networks, protein pathways, integrative cell function, and tissue as well as whole organ structure-function relations. Thus, these projects are not focused on only cellular signaling, but cell signaling is an important part of physiology and in near future the data collected and integrated by these projects might be possible to use to reconstruct intercellular signaling networks.

2.4.2 Reconstruction by spatial gene expression analysis

Beside the attempt of a complete integrated modeling of cellular signaling in humans, Diambra and da F. Costa (2005) present an example how intercellular signaling networks can be reconstructed from *Drosophila* data. The main purpose of their study is to improve the analysis of spatial gene expression patterns by means of complex networks. Images of small volumes of the organism show the gene expression intensities in a number of neighboring cells. An image is then transformed into a network of cells and two cell nodes are connected by an undirected edge if they have a similar expression intensity and are not further apart than a maximum distance. The basic assumption here is that cell signaling drives and coordinates gene expression at least in a local area. The analysis of the node degrees and clustering coefficients of the resulting networks could be used to characterize different stages in developmental dynamics and to identify abnormalities. Although this has been done for *Drosophila*, this approach can in principle be applied in any organism where images of gene expression intensities on the cellular level can be obtained. This shows how the

analysis of cellular signaling networks can be reasonably used to understand the function of organisms at a systemic level.

2.4.3 Reconstruction of nuclear receptor interactions

Considering human intercellular signaling again, the reconstruction of the signals from the available data as first step remains a problem. For this reason Albert *et al.* (2003) access the biomedical literature with an automated approach to generate a database of protein interactions with nuclear receptors. Therefore, a subset of MEDLINE texts is selected that contains terms from a dictionary (protein and nuclear receptor names as well as keywords like "bind" or "associate"). The dictionary is hierarchically organized (comparable to an ontology) and initially manually created, but subsequently extended by the achieved text mining results. The selected texts are decomposed into their sentences and that are then searched for co-occurring triples of protein, receptor and keyword terms. Finally stop lists containing rules that describe known false-positive results are applied and the resulting extracted interactions are stored in a database.

With this process, about 15 thousand co-occurrence triples were retrieved automatically from about 4 thousand abstracts. After manual curation of all results, about 3 thousand co-occurrences were classified as true-positive, which equals a precision (i.e. ratio of true-positives among all results) of about 20%. Interestingly, the number of detected interactions correlates with the number of published papers for a given receptor. Comparisons with yeast two-hybrid screen results suggest that such a correlation cannot be confirmed by experimental data. Thus, beside the problem of the uncertainty of automatically generated results from fuzzy natural language texts, it turns out that also text mining reflects the bias in the literature (see also the review on network extraction from text in Section 2.3.5).

This study shows how partial knowledge of intercellular signaling can be reconstructed from text. However, the locations (cell types or tissues) of the extracted protein and their receptors are not considered here. Thus, although the text mining approach is similar to the approach we will apply here (see Section 5 and the discussion in Section 5.5), the data gained by Albert *et al.* (2003) is not sufficient to reconstruct entire cell signaling networks.

2.4.4 Analysis of the human immune cell network

If a network of intercellular signals could be reconstructed, the next challenge is its analysis since such a network typically consists of a relatively low node number compared to a much larger number of connections. Especially the fact that any node pair might obtain a principally unlimited number of multiple edges (modeling the different first messenger relations between two cell types) is not considered in usual network or graph analysis. Therefore, Tieri *et al.* (2005) show how such a network can be analyzed by considering the number of different interactions as edge weight for shortest path calculations.

The network that Tieri *et al.* (2005) focus on is the human immune cell network, i.e. a subset of the whole intercellular communication network consisting of 19 cell types as nodes and a total of 316 connections, including autocrine self-loops. The data is taken from the

textbook based *Cytokine Reference Database* (Oppenheim *et al.*, 2000, not downloadable, but online available) and thus a curated subset of intercellular signals is used.

The high density of the network (the maximum number of edges is only $19^2 = 361$) shows that an analysis is only possible if the network is viewed from a different perspective. Therefore Tieri *et al.* (2005) propose measuring the *efficiency* of pathways between the nodes by taking the number of different connections between the cells into account. This is done by restricting each node pair to one edge at most per direction, combined with an edge weight that reflects the number of multiple edges. Now the weighted shortest path length between all node pairs can be calculated and models the distance between two cells.

The assumption here is that the more distinctive pathways between two cells, the more closely connected they are. Then, the efficiency of the whole network can be calculated by averaging the efficiencies of all node pairs. The influence of the mediators (i.e. first messengers in the immune system) that establish the communication between two cells is measured by comparing the efficiencies of the whole network including and excluding the mediator. The larger the drop in efficiency for a specific mediator, the more important is probably the respective substance.

From that approach, individual mediators can be compared regarding their importance on the communication efficiency of the network. Here it turned out that the most important mediators according to the network analysis are the same substances that have emerged in the last years as central components of the capability of the immune system. Thus, the results confirm the state of current knowledge. Additionally, the distribution of mediator relevances can be calculated. In this case it could be shown that a part of the relevance distribution shows a power-law behavior and thus, the mediators of the immune cell communication network might be connected in a scale-free manner.

2.5 Resulting implications

The objective of the present thesis is to reconstruct and analyze intercellular signaling networks in the human body. Cell signaling is understood here as the process of signal transmission between two cells, a source and a target cell. Through a ligand-receptor interaction, a ligand (or: first messenger molecule) is released from the source cell and transported to the target cell. The ligand finally binds to a complementary receptor at the surface or inside the target cell, initiating a signaling cascade inside the cell and causing cellular responses.

At the most detailed level, the nodes of the intercellular signaling network are the cell types sending and retrieving the first messengers. Depending on the data source the nodes might also represent entities of different anatomical levels, such as organs or tissues. Directed edges between the nodes represent the ligand-receptor interaction through which the respective connection is established. Multiple edges in the same direction are allowed since often more than one signal between two cells is existing.

The available data sources for network reconstruction are databases and the biomedical literature. For both kinds of sources it turned out that they do not contain explicit

information on complete cell signals. Instead, cell signals can be combined from partial knowledge, i.e. the sites of ligand and receptor molecules with known interaction relations are connected. A consequence from this general approach is that the resulting networks will mainly contain *potential* interactions or in other words, *hypotheses* about possible interactions.

A further consequence is that due to the combinatorial nature of such a reconstruction approach, the number of signals and thus the number of edges in the network might increase drastically. This is a challenge for visualization, validation and analysis of the reconstructed networks. Therefore, not only general cellular interaction models based on the biological components of a cell signal are developed, but also corresponding graph representations that allow a compact presentation of networks with high edge numbers.

Compact visualization supports the inspection and hence, the validation of the generated hypotheses. However, validation remains a problem since a number of hypotheses might be existing in reality, but have simply not yet been experimentally studied and reported. Additionally, a high number of edges compared to a relatively low number of nodes (there are approximately only about 200 different cell types in the human body, see Papin *et al.*, 2005) renders the analysis, i.e. the search for structures in such networks is more difficult, because it is likely that nearly any node may be connected with any other node.

For these reasons as first step a pilot study with the database CSNDB is conducted since it is one of the few structured sources that contain locations of ligand and receptor molecules at all. With this database the initial presumptions on the outlined problems of reconstructed cell-cell networks are examined. The development of a text mining approach is the next accomplished step, because already the initial inspection of the database showed that only few cell type locations are contained in CSNDB.

Both kind of data sources are finally applied to a specific task, the reconstruction of cell signaling especially between cell types relevant in neurodegenerative diseases. With such a restriction to a small set of cell types, the subset of hypotheses is small enough for more extensive validation and possibly providing new insights into the communication behavior of these cell types.

Chapter 3

Data structures for modeling cell-cell interactions

Contents

In this chapter, different ways to capture the biological process of intercellular signaling in models and graph representations are presented. First, three templates are introduced that model a single cell-cell signal with different granularity (Sec. 3.1). On the basis of these three biological models, three corresponding graph representations are developed (Sec. 3.2). This is done for two main reasons: to enable the use of partial information from databases or texts and to reduce the number of edges that would emerge if all signals were modeled separately. Fig. 3.1 summarizes the three templates and their corresponding graph representations.

3.1 Templates for single interactions

Intercellular signals can be represented by directly connecting two cells (column *1-comp*, upper part of Fig. 3.1). In this case, information on the messenger and its receptor that establish the connection is not modeled explicitly. The different entity types are considered here as *components*, therefore *1-comp* consists only of one component. Note that the cells on the left and the right side in Fig. 3.1 are not explicitly named as "source" and "target" cell respectively for simplifying the presentation.

However, this most straightforward representation has several disadvantages: first, any pair of cell types can be connected by a number of different signals (i.e., different messenger-receptor interactions). In the *comp-1* representation, the actual signals connecting the cells are not visible, but hidden in the properties of the interaction. Even more important, if

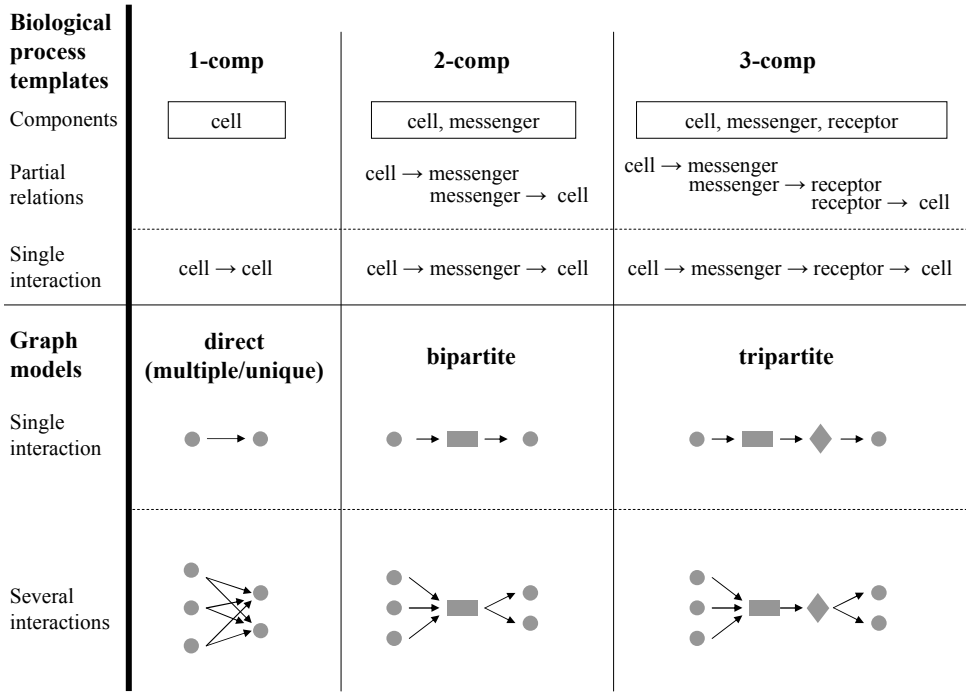

Figure 3.1: Schematic overview of different intercellular signaling representations. In the upper part of the figure (Biological process templates) a single interaction is modeled in three different granularities, named after the number of the different components (cell, messenger, receptor) the templates consists of. The number of components also reflects the number of partial relations into which an interaction can be decomposed (e.g. *cell* → *messenger*). The lower section (Graph models) shows the corresponding graph representations for single as well as for several interactions (i.e., the last row shows exemplary how the number of edges is reduced when different cells are connected by the same messenger-receptor interaction).

all interactions between a pair of cells are considered separately the number of interactions might increase strongly, resulting in very dense graphs. Furthermore, information inferred from the combination of partial data is not modeled adequately by the template *comp-1*.

For these reasons, two further templates representing single intercellular signals in greater granularity are developed and serve also as base for corresponding graph models. These two templates split up a cellular interaction and consist of two and three components respectively (*2-comp* and *3-comp*, see also Fig. 3.1, upper part), by applying additionally the entities *messenger* and *receptor* as components.

The number of components corresponds also to the number of partial relations into which a signal representation can be decomposed. So, for example, the *2-comp* template consists of two components, cell and messenger, and can be decomposed into two sub-relations, *cell* → *messenger* and *messenger* → *cell*. That means, partial information available on these relations could be gathered from different sources and afterwards combined

into a (potential) cell-cell signal.

The *3-comp* template extends this furthermore by explicitly modeling the receptor to which a released messenger is able to bind (Fig. 3.1, upper part). In this template, all relevant biological entities are included, i.e. all constituting elements of a single interaction can be immediately seen. Also this representation allows greater flexibility as well as a reduction of links when in the corresponding graph model several interactions are combined (Sec. 3.2).

3.2 Graph representations to combine single interactions

For each single interaction template (upper part of Fig. 3.1) a corresponding graph representation can be specified (lower part of Fig. 3.1). A graph $G = (V, E)$ consists of nodes $v_i \in V$ modeling the cells and edges $e_{ij} \in E$ connecting a node pair (v_i, v_j). In all graphs used here the edges are directed, i.e. if no multiple edges are allowed (see below) for any node pair two edges can be defined at maximum (one per direction).

Starting with the simplest graph model, the *1-comp* template can be straightforward converted into the *direct* graph model. "Direct" refers to the fact that the two cells, each represented as individual nodes (circles in Fig. 3.1), are directly connected, i.e. without considering any other component explicitly.

The direct graph model for signaling interactions can be further divided into a direct *multiple* and a direct *unique* model, employing multiple or single combined links between the cell type nodes respectively. In the direct multiple model, all different interactions that might exist between the same pair of cell types are represented as separate edges, whereas in the direct unique model all different interactions of the same direction between a pair of cell types are collapsed into one edge. That means in the direct multiple graph the number of edges can be arbitrarily large, whereas in the unique model this number is restricted to n^2 edges at maximum (self-loops are allowed to allow autocrine signals). Each edge in the direct unique model can be additionally equipped with the number of the contained interactions and further information.

To translate the other two single interaction templates (*2-comp* and *3-comp* respectively) into a corresponding graph model, *r-partite graphs* are applied. A graph G is called r-partite if the set of nodes V can be divided into r partitions such that all node pairs satisfy the condition $(v_i, v_j) \notin E$ with $v_i, v_j \in V_k$ and $1 \leq k \leq r$. That means, a graph is called r-partite if the graph can be divided into r distinct partitions where the nodes inside a partition are not connected and edges exist only between nodes of different partitions (see Chapter 1.6 in Diestel, 2000, and Fig. 3.2 for examples of 3-partite graphs).

It follows that nodes belonging to distinct partitions of the graph can be seen as possessing different types. In the bipartite and tripartite models used here (*bi* and *tri* are used as prefix instead of 2- and 3-partite), these different node types are expressed by different

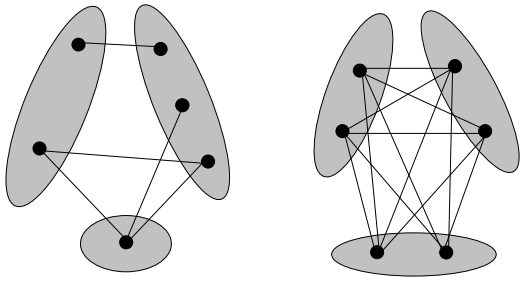

Figure 3.2: Two undirected 3-partite graphs as examples to demonstrate the definition of r-partite graphs. Each of the two graphs can be divided into three partitions. The nodes inside a partition are not connected, only edges between nodes of different partitions occur (figure adapted from Diestel, 2000).

symbols, i.e cells are still represented by circles, messengers and receptors as rectangles and diamond shapes respectively (Fig. 3.1, lower part).

Thus, by extending the set of node types in that way, both remaining templates for single cell-cell interactions (*2-comp* and *3-comp*) can be converted immediately into a corresponding graph representation. Beside the already mentioned advantages that an explicit modeling of messengers and receptors has (as e.g. for combining information from different data sources, see Sec. 3.1), another major benefit of these representations becomes visible when several interactions are combined into a network (Fig. 3.1, bottom section): any group of cells connected by the same messenger or messenger-receptor interaction can be combined into a bipartite representation in which the messenger (or the complete messenger-receptor interaction) is represented by a separate node. For a group of s source and t target cells that is completely connected by the same interaction, the number of edges decreases from $s \cdot t$ in the direct multiple graph representation to only $s + t$ in the bipartite model.

The addition of explicit receptor nodes in the tripartite model might further decrease the number of edges, though not as much as this is case for the transition from the direct to the bipartite representation. This further reduction is due to the fact that some messengers might share the same receptor. Then each messenger needs only to be linked to one receptor, which in turn contains all edges to the target cells, instead of linking each messenger to all target cells repeatedly.

Note that any single interaction can be converted into a bi- and tripartite representation. In the extreme case of only different messengers or messenger-receptor interactions nothing could be combined and the edge number would even increase, since any single interaction is then converted into a chain of two or three nodes (in the bi- or tripartite case respectively). But this can be neglected in our case as the results show (see Sec. 4.2.2 and Sec. 4.2.4).

The transformation between the three graph models is possible in both directions: mes-

senger and receptor nodes in the tripartite model can be combined into a single messenger node of the bipartite representation, which then can be resolved into all single interactions of the direct multiple model.

In bipartite and tripartite representations, however, adding and removing single interactions can not be performed easily. One possibility is to add or delete single interactions in the direct model and then to recalculate the bi- and tripartite representation. To perform the transformation in the other direction (from tripartite to direct representation) all interactions in the tripartite representation need to be stored separately. Then individual operations on single interactions can also be performed and the transformation process into the tripartite and bipartite model with combined messenger and receptor nodes performed afterwards. So, for any direction, if equal messenger or receptor nodes have been collapsed in order to reduce the amount of edges, operations on single interactions need recalculation.

Chapter 4

Reconstruction of cell-cell networks from CSNDB

Contents

This chapter presents how the Cell Signaling Networks Database (CSNDB) is applied to reconstruct cell-cell signals and to combine them into a network. As shown in Section 2.2, there are few databases available that contain relevant information that can be utilized. The CSNDB provides information on interactions of signaling molecules and their locations in the human body. This data can be assembled to reconstruct complete intercellular signals.

Therefore, in the following section the data scheme of the CSNDB is shown and how it is applied to the extraction of intercellular signals (Section 4.1). Subsequently, two

Class	Signal_Molecule	Cell_Signaling Gene_Expression	ExtraCell_Signaling
Fields	Endogenous/Exogenous	From_molecule	From_tissue
	Other_Name	To_molecule	To_tissue
	Is_Synonym	Interaction	Signal_Molecule
	Species		
	Type		
	Cell_Signaling		
	Tissue		
	Synthesis		
	Target		

Table 4.1: Class definitions in CSNDB. Only classes and fields used for the cell-cell signaling reconstruction are shown.

reconstruction approaches are then performed: in the first approach only information is accessed that can be directly detected as relevant, complemented by a second approach which is designed to exploit as much information from the CSNDB as possible. Both approaches and the resulting networks are shown in Section 4.2.

In Section 4.3 the subnetwork of organ-organ interactions resulting from the second reconstruction approach is used as an example how such networks can be used for further analysis. Section 4.4 briefly describes the implementation of the CSNDB extractions and finally, a discussion in section 4.5 closes this chapter.

4.1 Content and organization of CSNDB

The CSNDB (Igarashi and Kaminuma, 1997) is an object-oriented database designed according to the ACEDB format (a specialized genome database system developed by the Sanger Institute, see Walsh *et al.* (1998)) and applying the common Lisp language CLISP as inference engine. The contents of the CSNDB are generated by manual examination of biomedical research papers. Each entry in the CSNDB refers to the MEDLINE ID of the paper from which the information is taken.

Although the actual focus of the CSNDB is on *intra*cellular pathways, additional information is stored about molecules linking such pathways to extracellular signals. This information will be accessed by the extraction rules we present in this chapter. Therefore, the description of the CSNDB is mostly restricted to the data and structures relevant for *inter*cellular signaling.

The CSNDB was accessible through a web-interface (see Table 2.1 in Section 2.2) which allowed the user to search for specific molecules and interactions as well as to visualize selected pathways. For the purpose of the present work, a flat file of the CSNDB was downloaded (in July 2000) and used. Since then, no updates could be found. Currently

```
Signal_Molecule :  "GH"
   Endogenous
   Other_Name         "growth hormone"
   Other_Name         "hGH"
   Cell_Signaling     "GH-RH -> GH"
   Cell_Signaling     "somatostatin -> GH"
   Cell_Signaling     "GH -> IGF-1"
   Cell_Signaling     "GH -> GH receptor"
   Cell_Signaling     "GHS -> GH"
   Type               Hormone
   Tissue             "brain"
   Tissue             "Aorta"
   Tissue             "Placenta"
   Synthesis          "hypophysis"
```

```
Signal_Molecule :  "GH receptor"
   Endogenous
   Other_Name         "growth hormone receptor"
   Other_Name         "hGHbp"
   Cell_Signaling     "GH -> GH receptor"
   Type               Receptor
   Tissue             "brain"
   Tissue             "breast"
   Tissue             "heart"
```

```
Cell_Signaling :  "GH -> GH receptor"
   From_molecule      "GH"
   To_molecule        "GH receptor"
   Interaction        "ligand-receptor binding"
```

Table 4.2: Example for a definition of a signaling entity and its corresponding molecules in the CSNDB flat file. Only fields used for reconstructing cell-cell signals are shown here. Field values enclosed by double quotes references to other fields. References to other classes are realized by exact string matches of descriptors instead of identification numbers.

(i.e., autumn 2005) both the web-interface and the flat file are unavailable.

Several difficulties had to be solved in order to utilize the CSNDB. For example, the data is organized as mutually referencing objects, however, the flat file generated from the database contains only user-defined descriptor strings as identifiers. Since these descriptors can have typos or could otherwise be ambiguous, referenced objects often cannot not be identified. This is additionally worsened by the fact that some objects do not exist in the database. Such inconsistencies had to be resolved manually. Furthermore, the data structure definition is not in XML or in another format suitable for an automated data access. Thus, an automated processing of the flat file is not easily possible, and during the parsing process many other errors and problems had to be resolved.

4.1.1 Relevant classes

The CSNDB data structure consists of *classes* which contain *objects* implementing the class scheme. Therefore, each object consists of a name and a number of fields. The object names are used in the flat file to establish references between objects. So the fields of an object can contain values, references to other objects, or are used as boolean flags. In the latter case, the appearance of such a field means that its value is set to "true". References are enclosed by double quotes. A field might appear several times for different values or references (e.g., a molecule that appears in several different tissues) or is completely left out if no values are set (i.e., there are no empty fields). Table 4.1 shows the classes and fields mainly accessed in the present context to reconstruct intercellular signaling networks and Table 4.2 presents the objects `GH` (growth hormone) and `GH receptor` as typical example objects of the CSNDB.

Considering the relevant fields of a `Signal_Molecule`, such a molecule can be marked as `Endogenous` or `Exogenous`. Using this, exogenous molecules as pathogens, viruses or drugs can be excluded in the present context. Synonym molecule names are linked by the fields `Other_Name` and `Is_Synonym`. Sometimes the name of the `Species` containing this molecule is specified. Values of `type` might be e.g., `Hormone`, `Neurotransmitter` or `Cytokine`. A molecule can be assigned to more than one type. The field `Cell_Signaling` of a `Signal_Molecule` references all signaling interactions in which this signal molecule takes part.

The fields `Tissue`, `Synthesis` and `Target` are of special importance since they contain the names of the locations where the `Signal_Molecule` has been found, where it is produced or received, respectively. Here it turned out that although the field name `Tissue` suggests the use of a specific type of location (i.e. a *tissue*), this field can contain locations of very different kinds, as e.g. cell types, organs or organ systems, which are not all regarded as *tissue* in a biomedical sense and subsist on various levels of the anatomical hierarchy. Hence, in the following we prefer the term *location* (instead of tissue), which refers in the remainder of this chapter to entries in the fields `Tissue`, `Synthesis` and `Target`. In order to access the locations by their types, all locations finally used in the network are manually assigned to a location type (as e.g. cell type, tissue or organ, see Section 4.3).

In a `Cell_Signaling` object the two interacting molecules are specified in the fields

From_molecule and To_molecule and the type of the interaction is defined in the field Interaction. The type can be e.g., phosphorylation, protein-protein interaction or ligand-receptor binding.

Molecular interactions are also stored as Gene_expression, a class similar to Cell_Signaling, i.e. a Gene_expression object possesses all features of Cell_Signaling. Gene_expression is additionally considered here in order to capture events from steroid signaling where hormones bind to a receptor inside the cell and influence gene expression directly (Section 2.1). Further information about locations linked by intercellular signals is explicitly stored in ExtraCell_Signaling objects where two locations (in From_tissue and To_tissue) are directly connected through a Signal_Molecule. In some cases this information is also captured by the information in Cell_Signaling and its respective signaling molecules. Since the number of ExtraCell_Signaling objects in the CSNDB is considerably low, most reconstructed signals are inferred from interacting molecules and their locations.

4.1.2 Assembly of intercellular signals

Finally it has to be derived how intercellular signals can be extracted from the presented data scheme: from a Cell_Signaling objects the nodes and the links of the network can be inferred by connecting the locations of the interacting molecules From_molecule and To_molecule. With this information the templates that model a cell-cell signal (Section 3.1) are filled with the signaling molecules (ligand and receptor) and their locations. Hence, the locations as the nodes of the network include in case of the CSNDB reconstructions not only cell types, but also, for instance, tissues and organs.

As an example consider the growth hormone signaling GH -> GH receptor in Table 4.2: here the four locations of the GH molecule (brain, Aorta, Placenta and hypophysis) can be connected to three locations of the GH receptor (brain, breast and heart).

Thus, the directions of the links between the locations are determined by the From_molecule as source and the To_molecule as target nodes of the Cell_Signaling. The different semantics of the location fields (Tissue, Synthesis or Target) is in one reconstruction approach considered more specifically. Some fields of the Cell_Signaling and Signal_Molecule classes (as e.g. Endogenous/ Exogenous and Species, see Table 4.1) are used for filtering purposes.

Although the CSNDB flat file contains even more information about the molecules and their signals, the cell signaling reconstruction tasks use only the fields described here. Since most of the additional information does not appear very frequently in the selected molecules and signals, we would expect few changes in the accuracy or topology of extracted networks given further information. The two extraction runs on the CSNDB which are described in the following Sections 4.2.1 and 4.2.3 differ mainly in the selection of appropriate Cell_Signaling classes and in the handling of the different location fields.

Entity	CSNDB	Reconstruction I		Reconstruction II	
		Selected database objects			
		Total	Locations	Total	Locations
Cell_Signaling	1 382	169	74	180	106
Signal_Molecule	3 512	264	120	262	160
Gene_Expression	83	-	-	0	0
ExtraCell_Signaling	15	-	-	8	8

Locations		CSNDB	Resulting graphs							
			direct				direct			
			mult	uniq	bip	trip	mult	uniq	bip	trip
All	Nodes	172	85		159	205	94		215	287
	Edges	-	3 584	1 614	1 069	935	3 214	1 551	1 222	1 102
Organs	Nodes	88	29		74	107	57		155	213
	Edges	-	1 243	430	481	451	2 117	871	884	831

Table 4.3: Summary of CSNDB extraction results. The upper part of the table shows for the relevant database entities the number of appearance in the **CSNDB** and the numbers selected by the two applied reconstruction approaches. For both approaches the total number of selected entities (**Total**) and the number of entities for which locations are specified (**Locations**). The entities with locations could be used in the subsequent graph constructions. Note that the classes `Gene_Expression` and `ExtraCell_Signaling` are not accessed in the first reconstruction approach. The dimensions of the graphs resulting from the two approaches are shown in the lower part of the table (for all locations as well as for the subset of organ locations). The **CSNDB** column shows the total number of available locations and organs, whereas the other columns in the lower part of the table contain the node and edge numbers resulting for each available graph representation: direct multiple (**mult**) and unique (**uniq**), bipartite (**bip**) and tripartite (**trip**).

4.2 Reconstruction approaches and results

In this section two complementing reconstruction approaches applied on the CSNDB database are presented. In the first one (Section 4.2.1) only binary interactions that are explicitly typed as `ligand-receptor binding` are accessed. With this first approach it is tested how many and which kind of intercellular signals can be extracted from the CSNDB by the most simple filter. The second task (Section 4.2.3) is designed to exploit as much information from the CSNDB as possible, i.e. any type of `Cell_Signaling` is accessed and filtered for relevant molecules and their locations. Therefore, new extraction rules are designed in order to access relevant data that is missed by the previous approach.

Both approaches access the 3512 different molecules and 1382 signaling interactions contained in the CSNDB (Table 4.3):

$$M := \{\text{all } \texttt{Signal\_Molecule} \text{ entries}\}, |M| = 3512 \tag{4.1}$$

$$S := \{\text{all } \texttt{Cell\_Signaling} \text{ entries}\}, |S| = 1382. \tag{4.2}$$

Regardless of the chosen reconstruction approach, the general form of a `Cell_Signaling` $s \in S$ in the CSNDB is

$$s : m_1 + ... + m_k \rightarrow m_l + ... + m_n, \tag{4.3}$$

with $m_i \in M$ and $i \in [1...n]$. Further the conditions $k \geq 0, l = k + 1, n > k$ and $n \geq 2$ are met which means that the CSNDB signals consist of at least two molecules and a non-empty right side. The left side is in some cases empty. Some signals in the CSNDB contain an equilibrium symbol (`<->`) instead of the right arrow. This is not explicitly covered by Definition 4.3, but such signals are treated in the same way as the others.

Only signals that define locations for the relevant signaling molecules can be used. Since some of the molecules in the selected signals have no locations assigned, not all relevant signals are applied and thus, not all of the 172 available locations (Table 4.3) could finally be connected to other locations. Table 7.1 in the appendix (Section B) lists the locations for which connections could be inferred in either of both reconstruction approaches.

Both of the reconstruction approaches described in the remainder of this chapter select a subset of relevant `Cell_Signaling` interactions, determine two molecules of the signaling that contain the source and target locations and define how the locations are to be connected. At this point it is important to remember that the resulting links between the locations are *potential* intercellular interactions inferred from partial information in the CSNDB. That means, the information on the molecules and their interactions is taken from a validated database and the construction of the intercellular pathways follows biological plausible rules, but the resulting extracellular signals are still hypotheses.

4.2.1 Reconstruction I: Accession of binary ligand-receptor interactions

The first reconstruction approach applied on the CSNDB is the most simple one. Only binary ligand-receptor signals are selected since they certainly contain relevant cell signaling information. For this purpose, from all signals $s \in S$ only those possessing the form

$$s : m_1 \rightarrow m_2, \text{ with } m_1, m_2 \in M \tag{4.4}$$

are retrieved and checked for the following conditions:

$$\begin{aligned} &(1) \quad s.\texttt{interaction} = \texttt{"ligand-receptor binding"} \\ &(2) \quad m_2.\texttt{type} \qquad = \texttt{Receptor}. \end{aligned}$$

Further, the `species` field of both molecules should either be `Human` or missing. Since most molecules lack of a species entry and only a very few are marked as `Human` we decided to consider all molecules which are not explicitly assigned to non-mammalian species. Note that in the first reconstruction approach only `Cell_Signaling` objects are accessed, not `Gene_Expression` or `ExtraCell_Signaling`.

The last step in the reconstruction process is to draw the links between the locations of the ligand and receptor molecules. In this first and simple reconstruction approach the `Target` field of a `Signal_Molecule` is omitted and only the fields `Tissue` and the `Synthesis` are considered. Thus, all ligand molecule locations defined either in `Tissue` or in `Synthesis` are connected with all locations of the same fields in the receptor molecule.

4.2.2 Reconstruction I: Resulting interactions and networks

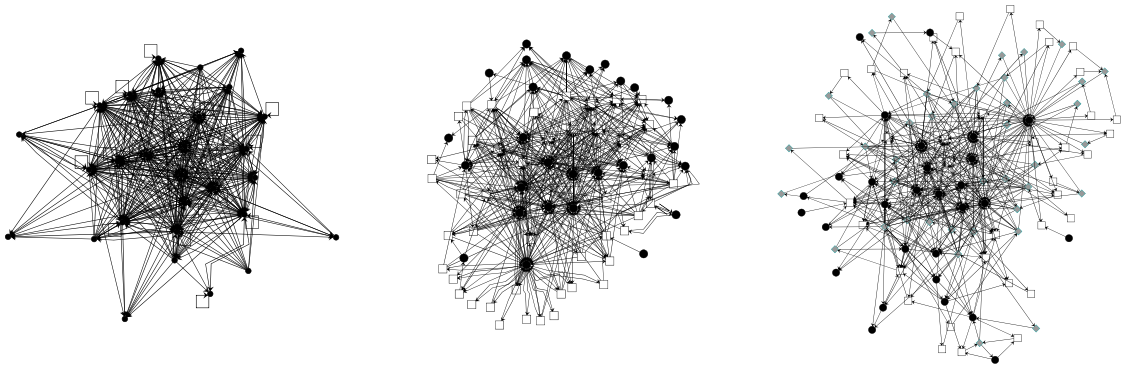

Figure 4.1: Different cell-cell interaction network visualizations resulting from the simple CSNDB reconstruction approach. Here only the unlabeled subset of the 29 organ locations is shown in the direct unique (left), bipartite (middle) and and tripartite (right) graph model (see also Table 4.3). All models possess a significantly lower edge number (about 430 to 480) compared to the simple direct multiple graph (about 1 200), which is not shown here. The locations are black circles, the ligand-receptor nodes in the bipartite graph are white boxes and in the tripartite graph the ligands are white boxes, whereas the receptors are shaped as gray diamonds. All figures are created using yEd from the yFiles library (`www.yworks.com`).

By applying the filter rules defined in the previous section on the CSNDB data, 169 signals consisting of 264 different molecules are selected (Table 4.3). If only the signals that contain locations for all relevant molecules are taken into account, 74 signals with a total of 120 different molecules remain which connect 85 locations (a complete list of the 74 remaining signals can be found in the appendix, Section B, Table 7.2).

From these selected signals and molecules, the graphs are constructed by adopting the different available models (Section 3.2). In the simplest case (direct multiple model) the resulting graph consists of 3 584 edges. The edge number can be reduced to 1 614 if only one edge is drawn at maximum between each node pair (direct unique model). Further reductions are obtained by using the bipartite and the tripartite graph model (1 069 and 935 edges respectively). Thus, the original number of edges can be reduced in some cases down to less than a third of the original amount, although the number of nodes increases due to the additional nodes required. However, in this reconstruction approach the connectivity of the locations, stays the same for all three different graph representations, i.e. locations linked in the direct model stay linked and no additional links emerge.

The edge reduction mainly has an impact on the visualization of the resulting graphs. Figure 4.1 shows the organ subnetwork of the complete network extracted from the CSNDB, comprising of 29 organ nodes. The organ subnetwork is chosen since it is the largest one in the CSNDB with entities at the same anatomical level. For the same reason the organ network is used for the sample application that will be presented below (Section 4.3). Although the different representations of the organ subnetwork in Figure 4.1 (direct unique, bipartite and tripartite from left to right) have similar edge numbers, they all show a great reduction compared to the simple direct multiple model (Table 4.3).

Cell_Signaling :	"renin -> angiotensin II"
Cell_Signaling :	"angiotensin II -> aldosterone"
Signal_Molecule :	"renin"
Type	Hormone
Tissue	"kidney"
Tissue	"colon"
Synthesis	"glomerulus"
Target	"blood vessel"
Signal_Molecule :	"angiotensin II"
Type	Hormone
Tissue	"blood vessel"
Signal_Molecule :	"aldosterone"
Type	Hormone
Target	"kidney"

Table 4.4: The renin/angiotensin system as it is contained in the CSNDB as an example of a set of signals consisting only of ligands, but relevant in respect to intercellular interactions. Here only the different location fields and the molecule type are shown.

Not only visualization benefits from the conversion into bipartite and tripartite model, but the extracted potential signals could be for example listed completely, sorted by ligands or ligand-receptor interactions and would be easier to search and to review by biomedical experts.

4.2.3 Reconstruction II: Accession of any molecular interaction

The main problems resulting from the first extraction task (Section 4.2.1) are:

1. Not all interesting molecular signals in the CSNDB are binary or possess the type `ligand-receptor binding`.
2. The different location fields (`Tissue`, `Synthesis` and `Target`) are not considered appropriately.

In the following, these two problems are discussed. Additional extraction rules for signaling selection and location connection are presented and finally complemented by the definition of a template that unifies all different location interactions which are filtered and assembled from the CSNDB with the new rules.

Problem 1: Accession of any type of `Cell_Signaling`

The question here is mainly whether it is possible to access more information in the CSNDB when all signals of the general form (Definition 4.3) are considered. This might sound surprising since the number of potential interactions extracted by the first approach is already very large. However, manual examinations of the cell signals in the CSNDB showed

that the database contains information that is not selected with the first reconstruction approach.

An example of such a relevant cell signaling is `NGF + TrkA -> CREB`. Here the interesting ligand-receptor interaction is on the left side, with the nerve growth factor `NGF` binding to the receptor tyrosine kinase `TrkA`. This produces a transcription factor `CREB` (a cAMP response element-binding protein), which is of no further interest for our purposes. But the interaction of `NGF` and `TrkA` should be selected.

A further example of interesting database content that has not yet been detected is a part of the renin/angiotensin system that is covered by the CSNDB. The respective signals are `renin -> angiotensin II` and `angiotensin II -> aldosterone` (Table 4.4). Although both signals are binary, they could not be found with the previous approach, because all molecules are hormones, i.e. these signals are not ligand-receptor interactions, but rather "ligand-ligand" signals. The biologically most plausible explanation for signals containing two ligands is that there are "hidden" interactions or other processes in between, i.e. `renin` is not directly interacting with `angiotensin II`, but rather causes the secretion of this ligand. In fact, in this case the underlying physiological mechanism is that `renin` is being produced in the kidneys and cleaves a molecule `angiotensin I` from its precursor `Angiotensinogen` which is synthesized in the liver. `Angiotensin I` is then converted into `angiotensin II` by a angiotensin-converting enzyme. `Angiotensin II` in turn causes the release of `aldosterone` in the adrenal gland.

Similar complex mechanisms are involved in all other ligand-ligand interactions in the CSNDB. However, since all these molecules involved in this kind of signaling are of interest, the ligand-ligand signals should be considered as "left molecule causes the production/release of right molecule".

Following these observations, a new extraction rule should be more flexible in selecting the source and target molecules of the potential cell-cell interaction. For this purpose we define in addition to the sets of molecules (Definition 4.1) and signals (Definition 4.2) two sets of relevant *molecule types*:

$$\begin{aligned} L &:= \{\texttt{Hormone}, \texttt{Cytokine}, \texttt{Neurotransmitter}\} \\ R &:= \{\texttt{Receptor}, \texttt{Ion\_Channel}\}. \end{aligned} \tag{4.5}$$

Here, L contains all ligand molecule types of the CSNDB, whereas R consists of molecule types with a receptor function which are the actual `Receptor` type, but also `Ion_Channel`, because ion channels can also serve as receptor for extracellular messengers (which change the conductivity of the ion channel, see Section 2.1.4). Examples of relevant interactions with ion channels in the CSNDB are `estradiol -> Maxi-K channel` or `L-glutamate -> GluR5`.

Using the molecule type sets (Definition 4.5) the molecules m_i of a signaling s can be

assigned to different groups, depending on the molecule type:

$$\begin{aligned} M_{left} &:= \{m_i, ...\}, \text{with } m_i.\texttt{type} \in L \text{ and } 1 \leq i \leq k \\ M_{right} &:= \{m_i, ...\}, \text{with } m_i.\texttt{type} \in L \text{ and } l \leq i \leq n \\ M_{lig} &:= \{m_i, ...\}, \text{with } m_i.\texttt{type} \in L \text{ and } 1 \leq i \leq n \\ M_{rec} &:= \{m_i, ...\}, \text{with } m_i.\texttt{type} \in R \text{ and } 1 \leq i \leq n \end{aligned} \tag{4.6}$$

Thus, M_{left} and M_{right} contain only ligand molecules from the left or from the right side of the signaling, respectively. M_{lig} and M_{rec} cover both sides of a signaling s, but contain either only ligand or only receptor molecules. All sets can either be empty or contain any number of molecules.

Finally, using these molecule groups, a filter function $select : S \rightarrow \{0, 1\}$ can be defined that selects all signals s with a specific ligand-ligand or ligand-receptor combination:

$$select(s) \;:=\; \left[\, (|M_{left}| > 0) \wedge (|M_{right}| > 0) \,\right] \;\vee\; \left[\, (|M_{lig}| > 0) \wedge (|M_{rec}| > 0 \,\right] \tag{4.7}$$

This function filters two kind of signals: these with at least one ligand molecule on each side of the signaling or those with at least one ligand and one receptor molecule at any position of the signaling. Hence, the first part of this filter would find signals like `renin -> angiotensin II` since both molecules are ligands. The second part of the filter accesses signals like `NGF + TrkA -> CREB`, because `NGF` and `TrkA` match the ligand and receptor condition.

After selecting a relevant `Cell_Signaling` object, two molecules with the respective source and the target locations to be connected have to be chosen, because the filter rule in Definition 4.7 might select a signal with more than two ligands or receptors. Manual examination of all selected signals revealed that in any case the molecule that provides the source locations is the first ligand and the target location molecule is either the second ligand or the receptor molecule.

Problem 2: Assignment of appropriate locations

The different location fields (`Tissue`, `Synthesis` and `Target`) have so far been treated uniformly, i.e., all locations in the fields `Tissue` and `Synthesis` of a ligand molecule are connected to all locations of the same fields in a receptor molecule. The `Target` field is completely omitted. A more accurate consideration should enhance the quality of the extracted paths, because synthesis and target locations are explicitly defined and thus, probably too many connections are drawn with the previous approach.

To illustrate this, consider again the Renin/Angiotensin system mentioned in problem 1 (Table 4.4). The molecule `renin` of the first signaling is located in the `Tissue` fields `kidney` and `colon` as well as in the `glomerulus` as `Synthesis` location. The `Target` location is `blood vessel`. The `angiotensin II` molecule of both interactions is reported as being located in the `blood vessel` as the only `Tissue` (no `Synthesis` and `Target` is given here) and `aldosterone` has a `Target` location in the `kidney`.

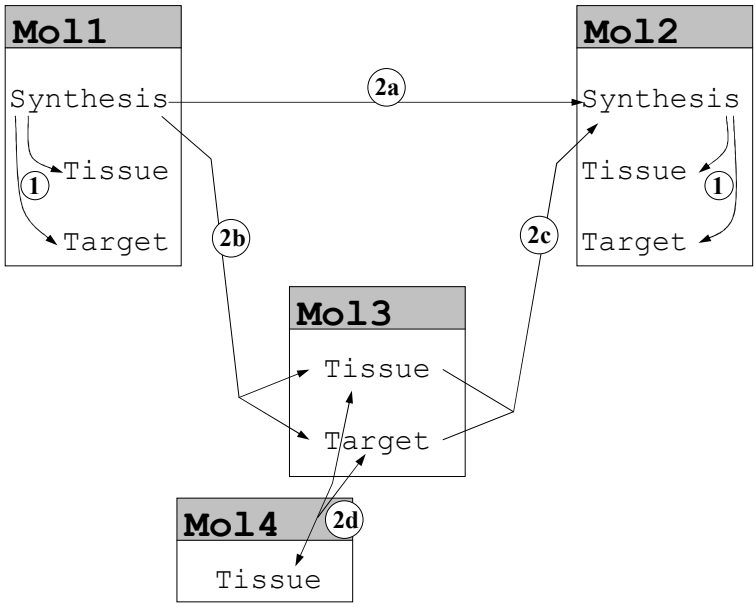

Figure 4.2: Schematic overview of the location connection rules in the second CSNDB reconstruction approach. The locations defined in different location fields of four abstract molecules (`Mol1` to `Mol4`) are connected according the new rules. The rule number that applies for each connection is denoted with encircled numbers at the arrows. The arrows mean that all locations of the location field at the start molecule are connected with all locations of the other location field at the end molecule.

Here it seems inappropriate to connect the `Tissue` *and* the `Synthesis` locations of `renin` equally to the `Tissue` location of `angiotensin II` since then there would be no difference between a synthesis location and other locations. The `Synthesis` location should rather be considered as the source and the other locations as targets.

Furthermore, since `angiotensin II` is contained in both signals, the previous assumption is supported, that such "ligand-ligand" interactions mean the induction of further messenger substances (as supposed in problem 1), i.e. `renin` is not binding to `angiotensin II` directly, but rather `renin` induces (by a mechanism not contained in the CSDNB) the production of `angiotensin II` which in turn amplifies the production of `aldosterone`.

Therefore, new rules for drawing links between the locations of CSNDB molecules are defined that reflect the biological reality more appropriate. These rules can be divided into two parts. First, there is a rule for connecting locations defined *inside* a single molecule:

1. **If** a molecule provides at least one `Synthesis` location and at least one `Tissue` or `Target` location,
   then connect the `Synthesis` locations with all `Tissue` and `Target` locations of this molecule.

Second, there are four rules defining the links *between* the locations of two molecules, depending on which location fields they provide (ligand and receptor molecules are here also denoted as *start* and *end* molecules):

2a. **If** both molecules provide one or more `Synthesis` locations,
then only the `Synthesis` locations of the start molecule are connected to all `Synthesis` locations of the end molecule.

2b. **If** only the start molecule provides one or more `Synthesis` locations,
then only the `Synthesis` locations of the start molecule are connected with all locations in the end molecule.

2c. **If** only the end molecule provides one or more `Synthesis` locations,
then all locations in the start molecule are connected only to the `Synthesis` locations of the end molecule.

2d. **If** there is no `Synthesis` location in any of the molecules,
then all locations of the start molecule are connected with all locations of the the end molecule.

Figure 4.2 shows these rules in a schematic view: consider four molecules, `Mol1` to `Mol4`, that contain different location fields, where `Mol1` and `Mol2` contain locations of all three location fields, `Mol3` and `Mol4` instead miss some location fields. For `Mol1` and `Mol2` first rule 1 can be applied, i.e. all locations contained in each molecule are connected from `Synthesis` to the other location types. Between `Mol1` and `Mol2` then only the respective `Synthesis` locations have to be linked (rule 2a). The rules 2b and 2c are applied between `Mol1` and `Mol3`, and between `Mol3` and `Mol2` respectively since `Mol3` does not contain any `Synthesis` location. Last, if none of the molecules comprises a `Synthesis` field, then the locations are connected similar to as it was performed in the previous reconstruction approach. This is applied for the location links inferred for `Mol3` and `Mol4`.

Regarding the quality of the links reconstructed by these rules it can be assumed that entries in `Synthesis` and `Target` fields are more meaningful than entries in `Tissue` fields. The reason is that entries in `Synthesis` and `Target` fields have a more specific semantic (secretion and binding), whereas the appearance of a `Tissue` field in a `Signal_Molecule` object means only that the molecule has been "somehow" observed in the respective location. Whether this substance is there since it is e.g. produced or consumed at this site is not specified. Also there is no documentation explaining the semantic of the CSNDB fields. Thus, the connection of synthesis and target location provides probably more certain information on intercellular signaling. However, in this context the main goal is to extract as many potential signals as possible and hence, also the connections between `Tissue` fields are considered.

Definition of a template for all extracted location interactions

The last problem to be solved arises through the application of the new rules defined above in this section: there is now no simple way to generate bipartite and tripartite

graph representations from the direct multiple model. In the previous reconstruction approach this was ensured by the fact that all locations specified in the molecule objects of a `ligand-receptor binding` are connected completely and thus can be easily combined with a new node in the bipartite model. In the second reconstruction approach however, each group of locations associated by a `Cell_Signaling` is not necessarily fully connected, because the different location fields (`Tissue`, `Synthesis` and `Target`) are now treated differently (see the description and solution of problem 2).

A further aspect that has not been considered yet is that it should be possible to remove single location interactions if they are negatively evaluated. Such deletions of individual interactions are not easy possible in a bipartite or tripartite model without affecting other interactions.

A way to solve these problems is to start at the other side and to create for each single location interaction an individual tripartite link, i.e. two location nodes are connected by a ligand and a receptor node. Then all equal nodes can be merged and thus the same kind of tripartite graph is generated as it has originated in the first approach from the direct graph representation (Section 4.2.1).

Following these considerations, a more flexible way to combine the several objects of the second reconstruction approach (i.e., locations, ligand and receptor molecules) is needed, and therefore, a template to store different interactions between locations is defined:

$$\mathbf{Loc_{source}} \rightarrow \mathrm{Lig}_{\mathrm{source}} \rightarrow \mathrm{Rec}_{\mathrm{target}} \rightarrow \mathbf{Loc_{target}} \rightarrow \mathrm{Lig}_{\mathrm{target}}, \tag{4.8}$$

meaning that a source location $\mathbf{Loc_{source}}$ is connected to a target location $\mathbf{Loc_{target}}$ via a ligand receptor-interaction. The respective molecules $\mathrm{Lig}_{\mathrm{source}}$ and $\mathrm{Rec}_{\mathrm{target}}$ are filled with the available information. For example consider the abovementioned signaling `NGF + TrkA -> CREB`. Among others, `NGF` contains the `Tissue` field `lung` and `TrkA` the `Tissue` field `liver`. Thus, an exemplary location link obeying the template in Definition 4.8 is:

$$\mathbf{Loc_{lung}} \rightarrow \texttt{NGF} \rightarrow \texttt{TrkA} \rightarrow \mathbf{Loc_{liver}}.$$

At this point another problem turns out: not for any reconstructed interaction the receptor is known since some are "ligand-ligand" interactions (as described in the discussion problem 1), i.e. a ligand molecule at the start site induces the production of another ligand at the target site. For this purpose the template in Definition 4.8 contains the variable $\mathrm{Lig}_{\mathrm{target}}$. In case of a "ligand-ligand" interaction the receptor variable $\mathrm{Rec}_{\mathrm{target}}$ is filled with a label that marks this part of the interaction as *unknown*. The second ligand is then stored in the $\mathrm{Lig}_{\mathrm{target}}$ field of the template.

For example, one of the location links based on the 'ligand-ligand" interaction `renin -> angiotensin II` (Table 4.4) is then

$$\mathbf{Loc_{glomerulus}} \rightarrow \texttt{renin} \rightarrow \texttt{?R\_renin} \rightarrow \mathbf{Loc_{blood\ vessel}} \rightarrow \texttt{angiotensin II}.$$

Thus, `?R_renin` is the label to denote that this value is currently missing. Similarly, missing ligands obey the form `?L_`*ligandname*.

All connections found between pairs of locations generate single location links that obey the template form shown in Definition 4.8. Finally all individual links are joined by combining all equal ligand and receptor nodes in the tripartite graph representation (except nodes representing unknown ligands or receptors). From this tripartite model, bipartite and direct representations can be generated easily.

Workflow

Summarizing the steps described above, the workflow of the second reconstruction approach is:

1. Selection of relevant `Cell_Signaling` and `Signaling_Molecule` objects
2. Connection of different location fields of the selected molecules
3. Generation of location-location interactions by applying a generic template
4. Merging of all equal nodes and creation of a tripartite graph
5. Generation of bipartite and direct representations

This second reconstruction process from the CSNDB was described for the class `Cell_Signaling`, but also the class `Gene_Expression` (Table 4.1) has been accessed with the same rules, since it is structurally similar to `Cell_Signaling`. As in the previous approach, all molecules with the field `Species` set to `Human` or missing are considered. Also unsuitable locations and impossible location pairs are ignored. Additionally, all molecular interactions that contain molecules with the flags `Exogenous` or `Endocrine_Disruptor` are ignored.

Supplementary to the molecular interaction classes, the class `ExtraCell_Signaling` is accessed. The information stored here can be directly translated into the template form (Definition 4.8) and does not need any further reconstruction. However, the number of `ExtraCell_Signaling` and `Gene_Expression` objects in the CSNDB is considerably low and thus, nearly all reconstructed signals are gained from accessing the `Cell_Signaling` objects.

4.2.4 Reconstruction II: Resulting interactions and networks

Applying the rules of the second reconstruction approach to the CSNDB results at first in the selection of 180 `Cell_Signaling` objects from which finally 106 can be used, because all of their respective 160 molecules (ligands and receptors) have at least one location entry (Table 4.3). Additionally, 8 of the 15 existing `ExtraCell_Signaling` objects contain information that is not covered by the `Cell_Signaling` objects and thus, these extracellular signals are also used. Unfortunately, no entry of the `Gene_Expression` fields matched the filter rules. Hence, information from `Gene_Expression` is not further used. Complete

lists of the 106 remaining cell signals and the 8 extracellular signals can be found in the appendix (Section B, Tables 7.3 and 7.4 respectively).

Some locations which are not suitable for our purposes are explicitly removed, as e.g. `pooled` or `cell line` as well as all locations related to ontogenesis, as e.g. `embryo` or `pluripotential stem cell` (a list with all extracted locations is given in Table 7.1 in the appendix, Section B). Also some location links are generally excluded, as e.g. such links that would connect male with female sex organs.

With the second reconstruction approach, 32 additional signals (compared to the previous approach) are found. Also the number of locations to be connected is higher. Interestingly, the number of links between these locations is lower (see the direct multiple or unique numbers for edges in Table 4.3). That is most likely due to the different connection rules between `Synthesis`, `Tissue` and `Target` locations.

The largest part of the selected signals are still binary ligand-receptor interactions. Most of the the newly acquired signals have no type defined and are found by the rule that allows "ligand-ligand" signals. Two of the new signals are protein-protein interactions and only a few of all relevant signals are not binary (for details see the complete list in Table 7.3 in the appendix, Section B).

From all selected signals, 3 214 links between all different 94 locations are inferred and checked for known interactions by biomedical experts, resulting in 452 reconstructed links that are known to exist (about 14.06%). It has also been found that 12 of the reconstructed links emerge from the use of a *target ligand* (i.e. the second ligand of a "ligand-ligand" signaling) as separate node. Another difference to the previous results is that the number of links between organs nearly doubled (for both direct representations as well as multiple and unique graphs, see Table 4.3). With three of the graph representations (direct unique, bipartite and tripartite) the edge number can be reduced by more than 50%. The organ subnetwork is shown in its tripartite representation in Figure 4.3.

Figure 4.3 shows not only a tripartite visualization as it has been shown for the first CSNDB reconstruction in Figure 4.1. Here, additionally, the strong components of the graph are calculated and are represented by the color of the nodes. In graph theory, a *component* of a graph G is a maximal connected subgraph of G, i.e. for any pair of nodes u and v in this subset there is a path from u to v (Diestel, 2000). Sometimes this is also referred to as *strong component* (Batagelj and Mrvar, 2003). A *giant strong component* (GSC) (or *giant component*) therefore is the largest strong component in G (Ma and Zeng, 2003).

The graph in Figure 4.3 is dominated by a large GSC (gray nodes) and several small components (black nodes) often consisting of only one location. In one case such a component is a completely separated subgraph (the *gastrointestinal tract* with the ligand *motilin* and its respective receptor in the upper right part of Figure 4.3). Many other components are single nodes, i.e. either source nodes (e.g. the *larynx* in the lower right part) or target nodes (e.g. the *tongue*, right above the larynx) with only edges into or from the GSC respectively. Some components are subgraphs consisting of more than one node (as e.g. the subgraph between *hippocampus* and *melatonin* at the lower left corner of the figure). Hence, the most important thing to note here is that the largest part of the network

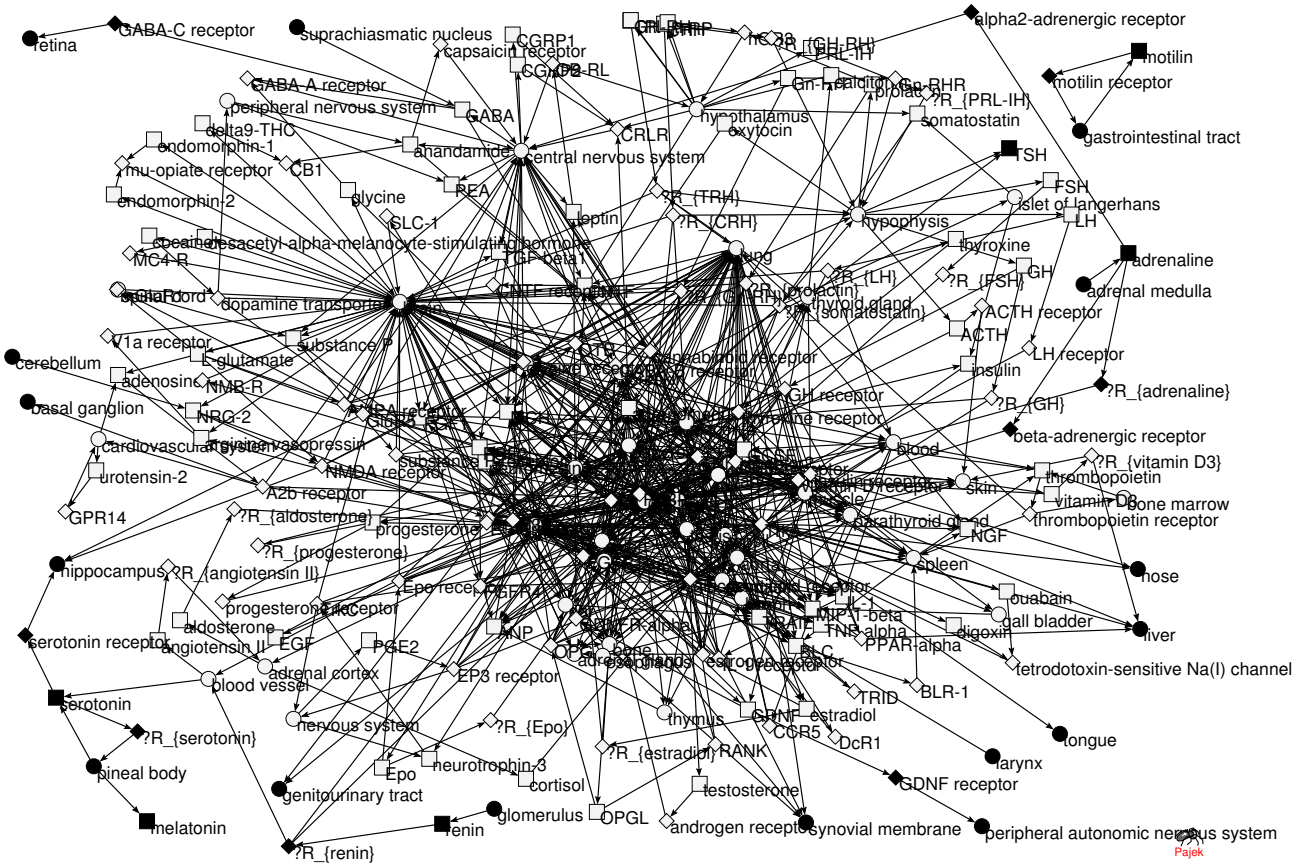

Figure 4.3: Tripartite organ graph resulting from the second CSNDB reconstruction approach. The color of the nodes indicates whether it belongs to the giant strong component (GSC, gray) or to any of the other components (black) of the graph. All nodes which do not belong to the GSC are placed around the GSC. The shape denotes (as in Figure 4.1) the locations (circles), the ligands (boxes) and the receptors (diamonds). The strong component calculation and the graph visualization are conducted with Pajek (Batagelj and Mrvar, 2003).

is connected in a way that each node can be reached from any other node.

To detect the actual density of the networks and other statistical measures, the direct unique graph representation should be used. Additional node sets and multiple edges unnecessarily complicate a quantitative analysis. Figure 4.4 shows the direct unique representation of the organ network derived from the tripartite (Figure 4.3) and bipartite (not shown) graph (Definition 4.2.3).

The density of a graph is calculated as the ratio of the number of edges e and the number of all possible edges: $density = \frac{e}{n*n}$, with n as the number of nodes. Here the number of possible edges is $n * n$ since edges in both directions and also self-loops are allowed. The organ graph has then a density of $871/52^2 \approx 0.268$. Further, the *diameter* of the graph, which is defined as the longest of all shortest pathways, is 4 for the organ network and the average distance among all reachable pairs is 1.764. These few numbers

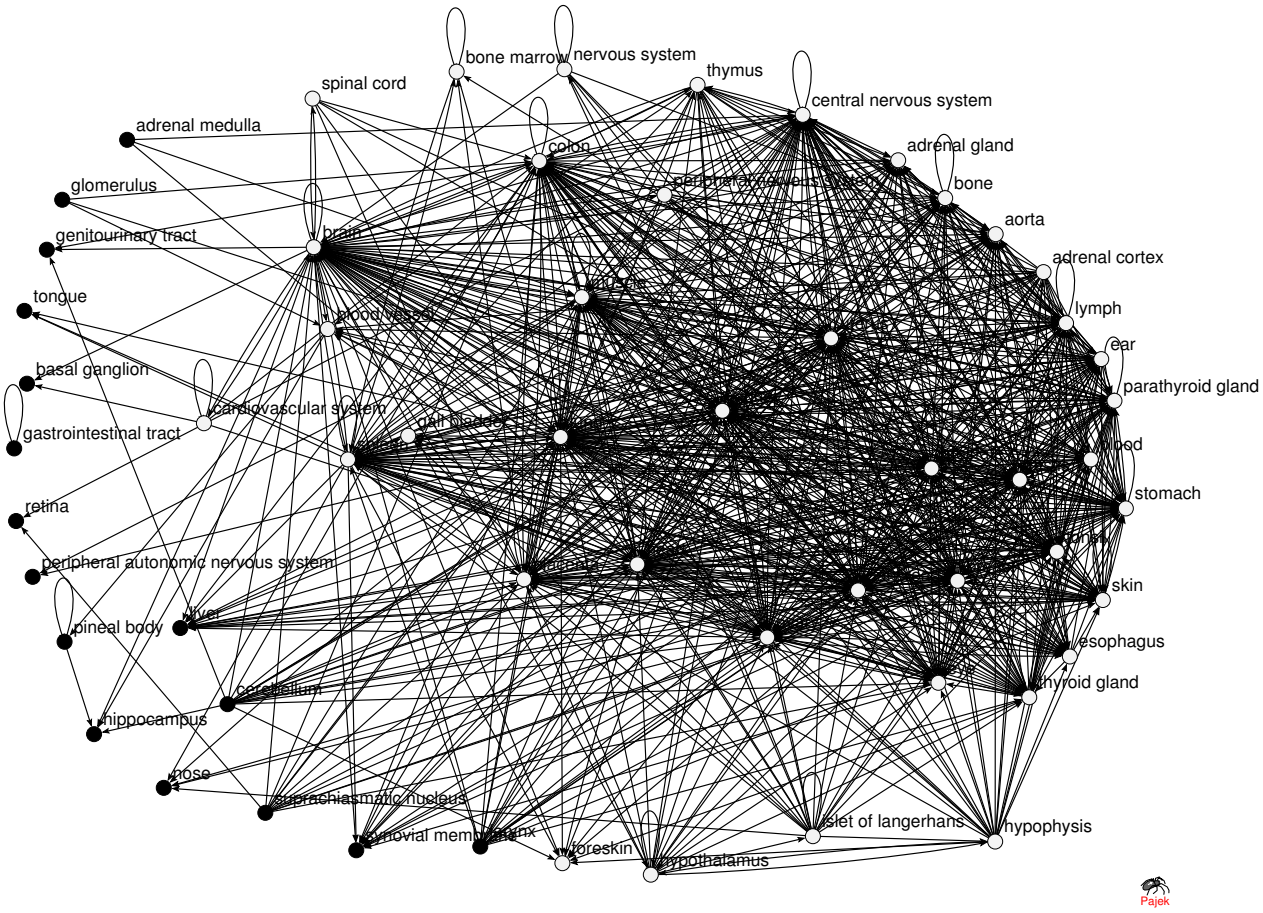

Figure 4.4: Direct unique organ graph resulting from the second CSNDB reconstruction approach. The nodes represent the organ locations, the node color indicates whether the node belongs to the giant strong component (GSC, gray) or to any of the other components (black) of the graph. All nodes which do not belong to the GSC are placed at the left and bottom side of the figure. The strong component calculation and the graph visualization are conducted with Pajek (Batagelj and Mrvar, 2003).

illustrate that not only the giant component connects nearly all nodes of the network, but that also the connectivity inside this component is very dense and that it is therefore hard to detect structurally different parts. The situation is similar for the complete network of all locations.

A further kind of statistical analysis is the distribution of node degrees. As it can be observed in Figure 4.4, there are nodes with only a few connections as well as nodes with a very high degree. A network is called *scale-free* if the distribution of the node degrees fits a power-law, i.e. decays linearly in a log-log plot (Barabási and Albert, 1999). That means the graph consists of many nodes with few connections and a small number of nodes with many connections, so-called *hubs*. This property makes the network scalable and more error-tolerant. In our case, the organ network consists of nodes with very different degrees, but no power-law could be found to fit this distribution (i.e., the respective correlation coefficient r is only $r \approx 0.49$, but a value close to 1 should be obtained). The same holds

for the complete network. Thus, although the node degrees show differences they did not fit to a known distribution.

4.3 Correlation of graph topology and biological behavior

It is a general problem of statistical network analysis in biology to relate graph topological measures to the biological reality that is modeled by the network. Metabolic networks of different organisms, for example, have been shown to be scale-free and to share other global properties under certain conditions (Jeong *et al.*, 2000). But it is still questionable what such global network properties can tell about the physical or biological behavior of the organism.

An example of such a relation is a study that shows correlations between the physical wiring lengths and network topology measures (as e.g. the average shortest path length) in cortical networks (Kaiser and Hilgetag, 2004). Inspired by this approach, we asked in the context of intercellular networks whether there may be a correlation between physical distances of organs in the human body and their signaling intensity, i.e. the number of different signals existing between a pair of organs. Here organs are considered since these are probably anatomical entities for which physical distance data might be found at all. Furthermore, the most nodes in the network we extracted from the CSNDB are organs.

A similar attempt to analyze especially dense intercellular signaling networks has been done by Tieri *et al.* (2005) for the human immune cell network (Section 2.4.4). In the following, the data resulting from the second CSNDB reconstruction approach (Section 4.2.3) is used.

4.3.1 Definition of distances

The goal of this sample study is to find out whether the signaling properties between organs differ depending on the anatomical distance between the organs. For example it might be that organs located close to each other do share more signaling interactions than far distant organs. For this purpose we need to define with which quantitative measures the signaling behavior in the network and the organ distance can be modeled. First, as a signaling property we define the *signaling intensity* between two locations loc_i and loc_j in a signaling graph G:

$$\text{signaling intensity}(\mathrm{loc}_i, \mathrm{loc}_j) := |E(\mathrm{loc}_i, \mathrm{loc}_j)|, E \in G, \tag{4.9}$$

which is simply the number of parallel edges of the same direction between a node pair in the direct multiple graph representations.

The signaling intensity is then compared to a physical distance measure of the two respective organs in the human body. Hence, the next question is how to define the physical distance of organs. One possible source of organ distances are models used in

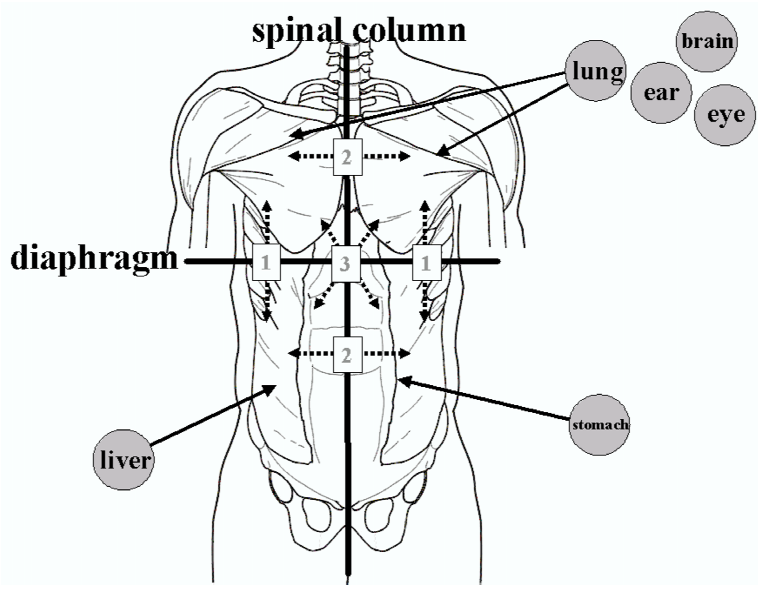

Figure 4.5: Body quadrants scheme. To measure distances between organs the torso is divided into four parts split by lines along the spinal column and the diaphragm. The organs are then assigned to one or several quadrants, depending on their location in the body. Organs not located on the torso (as e.g. the brain) are assigned to the upper two quadrants. The distances are finally chosen accordingly to the course of the blood stream (boxes on the axes).

nuclear medicine and created to measure the influence of radiation on different parts of the human body (Zankl *et al.*, 2003). These models are very accurate, distances are usually measured in millimeters, but, however, the data is not easy to obtain from the publications and, most importantly, does not cover all organs of our data set. So we decided to introduce another distance model, the *body quadrant model*.

Body quadrants are known from clinical practice. Here the torso is divided into four equally sized fields (split up horizontally by the diaphragm and vertically by the spinal column) and all organs can be assigned to one ore several of the quadrants (Figure 4.5). For example, the stomach is only located in the lower right quadrant, but the lung is contained in the upper two quadrants as well as all organs above the lung (brain, eyes etc.). Other organs like muscles or blood can also cover all four quadrants. This classification is done manually for all 57 organs in the CSNDB network.

Based on the body quadrant scheme, physical distances can now be defined according to the course of the blood stream (values on the axes in Figure 4.5): adjacent quadrants above and below the diaphragm (horizontal division) are closest and therefore organs located in such quadrants have a distance of 1, followed by the vertically adjacent quadrants which are set to a distance of 2. Finally, the diagonal distance between all quadrants is considered as 3. Organs of the same quadrant have a zero distance.

Since some organs are located in more than one quadrant, the definition of distances has to be extended. For example, the distance between the lung (located in both upper quadrants) and the liver (in the lower left corner) could be 1 if their closest junction is

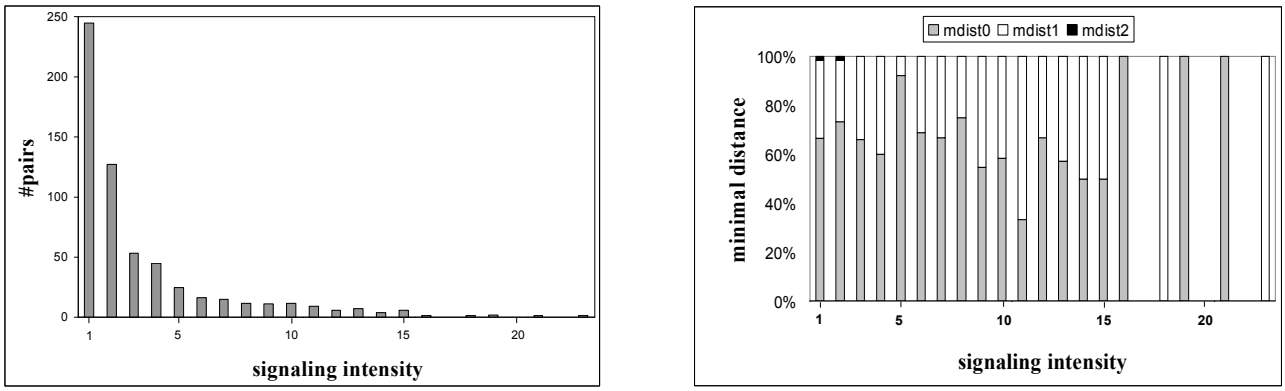

Figure 4.6: Results of the organ distance analysis. The **left panel** shows the distribution of the number of connections (i.e. signaling intensity) among the organ pairs. A bar in this plot represents the number of organ pairs connected by the respective number of edges. Each edge direction is therefore considered separately. On the **right panel** the results for the correlations of the signaling intensity with the *minimal distances* of organs are presented. Each bar shows the distribution of the three organ distances (0, 1 and 2; a distance of 3 was not observed) occurring among the organ pairs of a specific signaling intensity. Bars are left out if no organ pair exhibits this signaling intensity.

considered and 3 otherwise. Thus, three different distance measures are possible:

- *minimal distance*: consider only the minimal distance of all distance values
- *maximal distance*: consider only the maximal distance of all distance values
- *average distance*: calculate the average of all distance values

In case of the lung and the liver this would be 1 for the minimal distance, 3 for the maximal distance and 2 for the average distance.

4.3.2 Results

Figure 4.6 presents the results of this sample analysis. On the left side the distribution of the signaling intensity among the organ pairs in the reconstructed network is considered. Many pairs in the organ network are connected by only a few parallel signals, and the higher the signaling intensity is, the lower is the number of node pairs exhibiting this intensity. Thus, this decreasing distribution shows that there is a variety in the organ pairs in respect to the signaling intensity, which is a prerequisite for further investigations. If all node pairs would be connected by a similar numbers of links, then a search for correlation with other measures would make no sense.

On the right plot of Figure 4.6 the results for one of the distance measures, the minimal distance, is presented. From the four possible distance values (0, 1, 2 and 3) the distance 3

does not occur and a distance of 2 occurs only a few times for low signaling numbers. The two remaining distances are distributed in a nearly homogeneous way. There is no observed tendency from which to draw conclusions such as e.g., highly connected organs are located more closely than other organs. Even the organ pairs with the most connections (at the right end of the figure) show both distances.

The other distance measures were also tested. Additionally, averages per signaling intensity class are calculated and weighted by the size of the class. In all cases the results are similar and no clear correlation of signaling intensity and physical organ distance could be observed.

The reasons for this might be found on both sides. Both the abstract network measures and the distance measures are not based on biological reality. Physical distance might not correspond to the time on average a hormone takes to travel between one tissue and another. Also one must have in mind that this is only a network of hypotheses, i.e. of potential interactions, and that with a more accurate network the distribution of the signaling intensity could change. Furthermore, the resolution of the physical distance value is probably not high enough. Many organs share a quadrant or are at least directly related with distance 1. Hence, only two distance values mainly occur. Here it could be helpful to try to apply finer granulated organ distances from nuclear medicine (Zankl *et al.*, 2003), even if such measures exist for only a part of the organ network.

Thus, an interdependency of a topological graph measure and a physical property could not be shown, but however, the approach can be re-used and applied to more accurate data.

4.4 Implementation

The CSNDB is locally stored as an ASCII flat file in a proprietary format. This flat file is parsed and the results stored in a relational SQL database (PostgreSQL 7.2.1). Also all subsequent reconstruction and analysis processes access this database. The programming language is Java (Version 1.4.1), the database accession uses JDBC. As graph library the commercial yFiles library (`www.yworks.com`) is applied (Version 2.0.1.3). Graph analysis is conducted using Pajek 1.0.9 (Batagelj and Mrvar, 2003) and the R 1.9 statistics software (`www.r-project.org`) with the Bioconductor package (`www.bioconductor.org`).

4.5 Discussion

The approaches to reconstruct intercellular signaling from the Cell Signaling Database (CSNDB) can be seen as a sample for the extraction of this kind of signals from molecular databases in general. Although the CSNDB has a different focus compared to other databases (Section 2.2), all reviewed molecular databases have in common that they do not contain direct information on intercellular connections, but the signals have to be assembled from partial information. Thus, two problems remain for probably most database based reconstruction approaches: the molecules are often not properly assigned to a well-

defined anatomical location and even if this is the case, the combination of source and target locations into extracellular signaling pathways can easily result in great numbers of unvalidated hypotheses.

The location problem might be resolved by applying other data sources containing further information about the anatomical locations of the molecules. For this purpose special anatomical ontologies, such as FMA, eVOC or cytomer (see also at the end of Section 2.2), could be used to assign the type of locations (whether it is e.g., a cell type or an organ) and their positions in the anatomical hierarchy (whether it is e.g., "part-of" a tissue or an organ). Additionally, in case of the CSNDB the semantic of the location fields is uncertain. Also the criteria for the selection of the incorporated papers are not documented and hence it is hard to determine whether and in which direction the CSNDB content is biased and whether it maybe lacks important cell signaling knowledge. The fact that the CSNDB seems not to be supported anymore, the difficulties in parsing the flat file and the general lack of documentation add to the uncertainties with this database.

But even if the same reconstruction approach would be applied to more accurate databases, the second problem of combining partial information into large amounts of unvalidated hypotheses remains. Such hypotheses are not easy to evaluate since they could be true, but simply might have not been experimentally investigated yet. For the reconstructed networks and their quantitative analysis this means a great number of uncertain edges resulting in dense networks that are difficult to analyze.

To support visualization and manual examination, different graph representations are introduced that reduce the number of edges. It could be shown that a part of the reconstructed cell interactions are valid, even if the underlying molecular interactions are not explicitly labeled as relevant (e.g. "ligand-ligand" interactions in contrast to `ligand-receptor binding`s). For this purpose new extraction rules are introduced that access as much relevant information as possible. However, the extracted networks are still too dense and uncertain for detecting structures by statistical analysis. Restrictions of the data to a subset of valid interactions or to a set of locations interesting in a specific application are possible strategies to make use of the extracted networks.

Thus, the presented preliminary database study shows exemplary the problems occurring generally in intercellular signaling network reconstruction. Since all currently available databases lack appropriate molecular location information and consist of partial information, text mining was considered as a reasonable alternative to gain cell signaling networks. Here the desired cell type locations can be defined in advance and are not restricted to a specific database content. Although the resulting networks consist also in this case of hypotheses, the number of considered cell types can be adjusted in order to allow more extensive validations.

Chapter 5

Reconstruction of cell-cell networks from text

Contents

Following the preliminary database reconstruction study, in this section it will be described how a text mining approach is conducted in order to reconstruct intercellular interaction

networks. Therefore, the ONDEX framework for data integration and network extraction is developed as collaborative work and applied to cell relation mining (Section 5.1). An ONDEX database containing relevant texts and background information (as biomedical ontologies and databases) is created (Section 5.2) and used (Section 5.3). The chapter closes with a brief summary of the implementation (Section 5.4) and a concluding discussion of the results (Section 5.5).

5.1 ONDEX as text mining framework

ONDEX (Figure 5.1) means "ONtological inDEXing" and is a general framework to support *database integration*, *text mining* and *graph analysis* (Köhler *et al.*, 2004). In different tasks, ontologies and their basic data model serve as a tool for interpretation and unification of different data. Although ontologies are often very differently defined in literature, they can be generally regarded as data structures for storing knowledge by linking *concepts* with different types of *relations* (Section 2.2).

The background knowledge stored in ontologies is applied to support data integration and text mining on a *semantic* level. For this purpose, different databases and ontologies are imported into ONDEX and converted into a standardized *concept based* data structure. Then the concepts are aligned and indexed in the imported texts. This enables searches for concepts in annotated texts. Hence, terms are not only found by their matching characters in the texts, but their meaning as defined by the ONDEX ontology can be taken into account. In the following, this approach will be termed as *concept based* approach since all processes in the ONDEX system are based on concepts and relations gained from external sources.

A further advantage of applying ontologies for data integration and text mining is their *graph based* structure. Ontologies are usually implemented as *directed acyclic graphs* (DAG) with the concepts as nodes and the relations as directed edges. The direction and the type of an edge indicate the kind of relation.

In this sense, the graph based structure of ONDEX supports not only database integration and text mining, but in fact is able to reflect any kind of biologically meaningful interactions. Thus, *network extraction* using ONDEX can be achieved in an integrative manner: interactions of molecules or other entities can be extracted from text by applying existing knowledge about the relations of the considered entities.

The development of the ONDEX system is a joint work together with Jacob Köhler (Rothamsted Research, Harpenden, UK) and Alexander Rüegg (Bioinformatics and Medical Informatics Department, Bielefeld University, Germany) based on original ideas of Jacob Köhler and early implementations of student projects at Bielefeld University. My own contributions to ONDEX take place mainly in the text mining part (development and implementation of the indexing methods, including a scoring function for homonym detection, and of the actual information extraction methods). Additionally I developed parsers for several databases, ontologies and also for the MEDLINE text import. ONDEX will be further developed in the group of Jacob Köhler at Rothamsted Research. It has been also

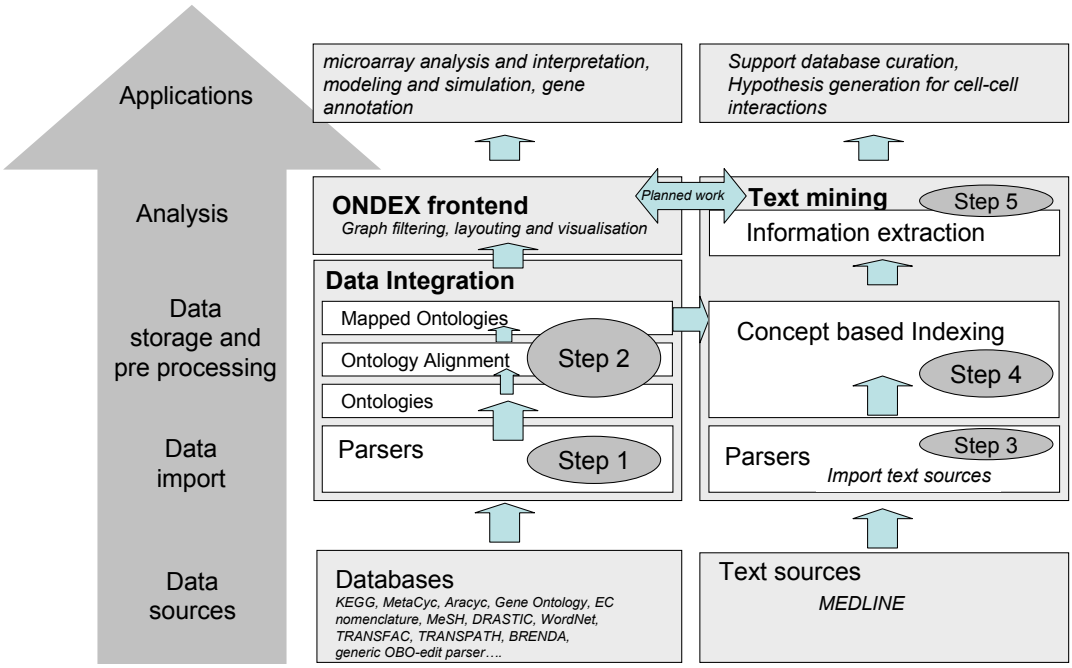

Figure 5.1: Schematic overview of the ONDEX system (adapted from Köhler *et al.* (2006)). The three main parts are: data integration, text mining and graph analysis (ONDEX frontend). Steps 1 to 5 are the parts of ONDEX used for extracting intercellular signaling from text.

extended towards graph analysis (ONDEX frontend, see Baumbach, 2005; Taubert, 2005) and applied to a text mining problem on plants and fungi (Winnenburg, 2005; Winnenburg *et al.*, 2006). Currently it is being re-designed with the main goal of a better end-user convenience and a greater modularity. The system is already available as open source software at Sourceforge (`http://sourceforge.net/projects/ondex`).

In order to extract intercellular signaling networks from text, five major steps are performed (Figure 5.1): appropriate databases and ontologies are selected and imported (step 1) and aligned (step 2); text sources are selected and imported (step 3) and indexed by a selected number of concepts (step 4); finally several information extraction processes (relation mining by concept based co-occurrence searches and hypotheses generation) are performed (step 5) to reconstruct extracellular signaling pathways.

In the following, after presenting basic definitions and the ONDEX database scheme (Section 5.1.1), the general function of ONDEX will be described according to these steps (Section 5.1.2 to Section 5.1.4), whereas the data integration part and the graph analysis module (ONDEX frontend) are described only briefly since the main focus is on text mining and network extraction.

5.1.1 Basic definitions and ONDEX framework overview

Ontologies

Although ontologies are defined in many different ways in literature (Gruber, 1993; Köhler *et al.*, 2003, 2004; Smith *et al.*, 2005), in the most basic and still widely used form an ontology can be described as an extension of a *controlled vocabulary* CV, which is a collection of well defined terms:

$$CV := \text{named set of concepts } c, \tag{5.1}$$

with a *concept* c defined as

$$c := (\text{identifier, names, definition}). \tag{5.2}$$

The symbols *identifier*, *names* and *definition* represent different data types and can be implemented in various ways. The *identifier* can be a string or a number, but must be unique, *definition* is usually a string and *names* a set of strings. A concept can consist of more than one name, i.e. usually a term and its synonyms are stored here. There can be a preferred or main term, but concepts are not identified by one of their names. This ensures that a CV does not contain ambiguous concepts such as e.g. *homonyms* (terms that obey the same name, but different definitions). A CV can still contain homonym names, but the respective concepts have different identifiers and an unambiguous definition.

Thus, a *concept* can be uniquely identified and hence, a controlled vocabulary is not only a loosely term collection. Each concept appears only once with one specific meaning.

An ontology O can then be defined as an extension of a controlled vocabulary (Köhler *et al.*, 2003):

$$O := \text{Graph } G(CV, E), \text{with edges } E \subseteq CV \times CV. \tag{5.3}$$

The types of the edges (i.e. the relation types) are given by a function t defined as:

$$t : E \rightarrow T, \text{with } T := \{\text{set of possible edge types}\}, \tag{5.4}$$

i.e. T describes the semantics of an edge in natural language and its algebraic relational properties (transitivity, symmetry and reflexivity).

All ontologies have an edge type 'is-a' $\in T$. If two concepts $c_1, c_2 \in CV$ are connected by an edge of this type, the natural language meaning is "c_1 is a c_2". For example, the concepts "vertebrate", "animal" and "organism" are connected by transitive 'is-a' relations, i.e. vertebrate 'is-a' animal and animal 'is-a' organism. The transitive 'is-a' relations can then be used to derive the fact that vertebrate 'is-a' organism. Furthermore, an ontology is defined as an acyclic graph in respect to its 'is-a' relations, i.e. circular definitions regarding the 'is-a' structure are not allowed. A further common relation type is 'part-of'. Examples for widely used ontologies are the Gene Ontology (GO, see Ashburner *et al.*, 2000), the Unified Medical Language System (UMLS, see Bodenreider, 2004) and WordNet (Fellbaum, 1998). Another ontology important in this context is the Cell Ontology (CL, see Bard *et al.*, 2005) that contains a hierarchy of cell types.

A similar concept is that of a *semantic network*, with the difference that in a semantic network the type of the relation between two concepts is not as strongly defined as in an ontology. Also in a semantic network it is not a requirement that the 'is-a' relation must exist. Thus, an ontology is a form of a semantic network, but not any semantic network can be regarded as ontology. In computational biology semantic networks are used for example to model intracellular signaling pathways (Hsing *et al.*, 2004). Sometimes the UMLS is also referred to as semantic network (McCray and Nelson, 1995).

ONDEX database scheme

Databases, ontologies and other sources are imported via specialized parsers into the ONDEX database by converting the data structure of the external source into the concept-relation scheme of ONDEX. Figure 5.2 shows the main parts of the entity-relationship diagram of the ONDEX database as far as they are concerned in this thesis. These parts are the ONDEX core (left side of Figure 5.2), storing the data sources as one unified ontology graph, and the text mining part (right side of Figure 5.2), containing selected texts and concepts to be mined as well as the text mining results. A further part of ONDEX is the *generalized data structure* (GDS) that allows to import data which has no pre-defined data types in the ONDEX database (not shown in Figure 5.2). This allows greater flexibility for importing and managing heterogeneous data, but is not used within this thesis.

The central table of the ONDEX core is `CONCEPT`. Here the identifier (field: `id`) and the description (field: `description`) of a concept (according to Definition 5.2) are stored. Each concept can consist of one or several names (synonyms) in the joined table `CONCEPT_NAME`. One of the names can be flagged as the preferred or main name (field: `is_preferred`), but this is not necessary for an unambiguous identification of concepts. So, synonyms are not stored as related concepts, but rather as properties of the concept itself. The field `is_unique` is set to `true` if there are no other concepts carrying the same name (homonyms) in the database. Since the concept names are processed with several natural language processing (NLP) tools (see following Section 5.1.2, step 1) there are several fields for storing the original name and different resulting names after the NLP processing. The table `CONCEPT_ACC` is used for storing different accession numbers of the same concept. An accession number in this context is a reference that maps a database entry to entities in other databases, i.e. a protein in Swiss-Prot for example can have links to corresponding genes the Gene Ontology.

In this version of ONDEX relations defined in the imported data sources are distinguished from mappings created by algorithms in ONDEX. For this purpose, imported relations are stored in `RELATION` and additional mappings between concepts derived by one of the ONDEX mapping algorithms (see Section 5.1.2, step 2) are written into `MAPPING`. Each relation and mapping is further characterized by a `RELATION_TYPE` (e.g. `is_a`) and a `MAPPING_METHOD` respectively. In a `RELATION` the direction is considered by discriminating between the source (`FROM_CONCEPT`) and the target (`TO_CONCEPT`) whereas a `MAPPING` consists simply of two concepts without any order. The re-designed ONDEX system currently under development unifies relations and mappings into the `RELATION` table by assigning

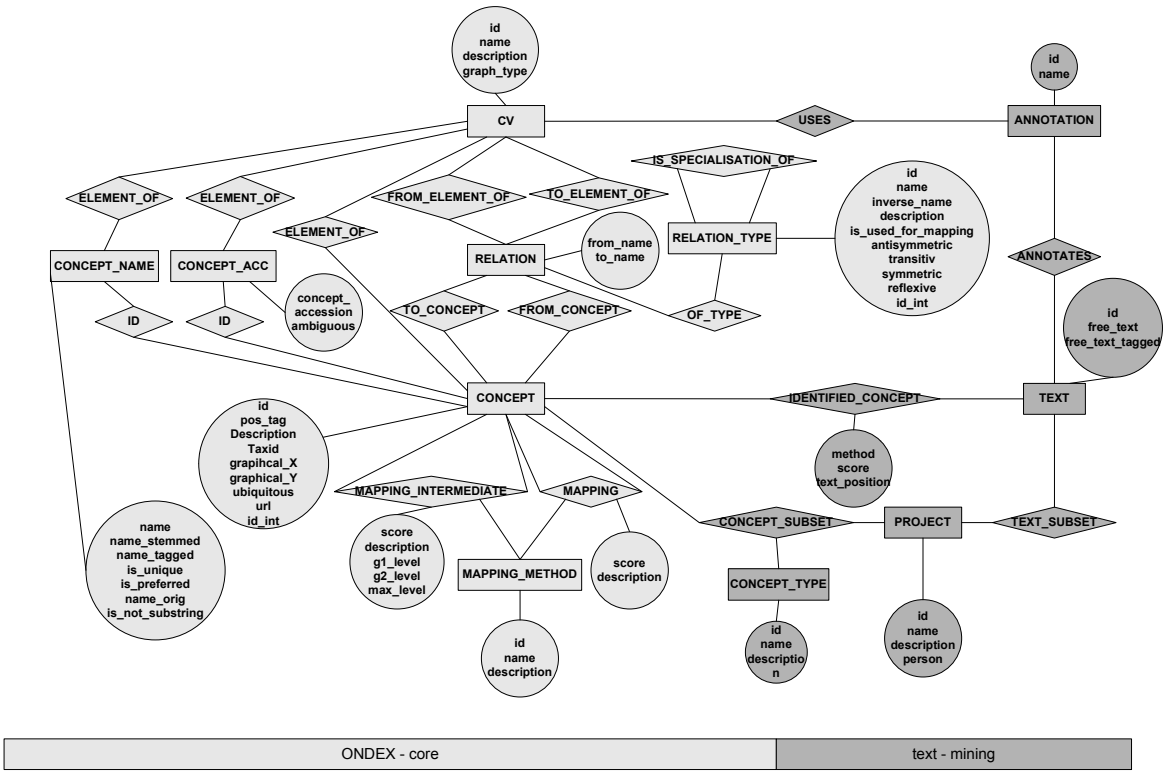

Figure 5.2: Entity relationship diagram of the ONDEX core and the text mining part.The generalized data structure (GDS) part is not used in this thesis and therefore are left out.

appropriate types in `RELATION_TYPE`. Undirected mappings will then be represented by double entries (one for each direction) in the `RELATION` table.

A further central table of the ONDEX core is `CV`. `CV` means *controlled vocabulary* and contains identifiers for all imported data sources, i.e. any concept and relation can be queried for its origin. Internally derived mappings are also identifiable by a `CV` entry.

The `CV` table also reflects our basic understanding of ontologies implemented in the ONDEX system: controlled vocabularies in the beginning contain concepts consisting of an identifier, a description and a set of names. The concepts can then subsequently be linked by relations (either imported or newly created) constituting altogether the ONDEX ontology. Thus, ONDEX is able to import simple controlled vocabularies without any pre-defined relations as well as ontologies and databases (at least those databases that can

CV	Full name	Content	Concepts	Relations
CL	Cell Ontology	Cell types	699	1 056
EC	Enzyme Nomenclature Committee	Enzyme classification	4 502	4 496
DRA	Drastic Insight Database	Host-pathogen interactions	5 438	9 907
TF	Biobase TransFac	Transcription factor database	13 945	9 179
GO	Gene Ontology	Gene and gene product attributes in any organism	17 427	25 337
AC	AraCyc	Biochemical pathway database for arabidopsis	18 909	18 424
MC	MetaCyc	Metabolic pathway database for different organisms	21 345	34 384
MESHD	Medical Subject Headings (MeSH) Descriptors	NLM's standard medical index terms	22 995	40 525
TP	Biobase Transpath	Signal transduction database	45 243	59 825
WN	WordNet	Lexical reference system of English nouns, verbs, adjectives and adverbs	115 424	213 335
TX	NCBI Taxonomy	Organism names and classification	220 927	220 927
BREND	BRENDA database	Enzyme information system	286 983	691 799
KEGG	KEGG database	Kyoto encyclopedia of genes and genomes	2 170 188	3 073 502

Table 5.1: Overview of some of the ontologies and databases available for importing into ONDEX. The gray lines mark the data sources imported and used within this thesis. "CV" is the identifier of the database/ontology. The table is ordered by increasing numbers of concepts and relations provided by the data source.

be interpreted as ontologies). The unified concept-relation structure supports the search for additional mappings between concepts of the various imported data sources.

The text mining part of ONDEX consists mainly of the texts imported into the table `TEXT`. There they are stored unchanged as well as processed by the same NLP tools used for importing the databases and ontologies. In case of mining huge amounts of text, a split into a number of text tables might be necessary (see step 3 in Section 5.1.3 for a more detailed problem description and Section 5.2 for the specific procedure applied to the MEDLINE texts imported in this context).

For each text mining task a `PROJECT` can be defined that specifies a subset of selected concepts (`CONCEPT_SUBSET`) and a subset of selected texts (`TEXT_SUBSET`). The concept based indexing algorithm (see Section 5.1.3, step 4) then assigns in `IDENTIFIED_CONCEPT` to each concept of `CONCEPT_SUBSET` the texts where it is found.

With the table `ANNOTATION`, texts can be additionally annotated by entries of other controlled vocabularies. The MEDLINE database for instance provides links to databases and other external sources.

5.1.2 Data integration

Step 1: Import of ontologies and databases

Each data source to be used in ONDEX (e.g. databases, ontologies, taxonomies etc.) requires the implementation of a specific parser that converts structure of the data source into the ONDEX database scheme. In ONDEX any data sources can be imported that obeys the form of a controlled vocabulary (Definition 5.1).

The parsers apply natural language processing (NLP) tools to the concept names. These are a word stemmer to restrict each word of the name to its linguistically main part and a part-of-speech tagger to add to each word its grammatical function (for a description of the specific tools used refer to Section C in the appendix). The results are used for the mapping and the indexing algorithms (see next step (step 2) in this section and step 4 in Section 5.1.3, respectively).

All parsers currently available in ONDEX are listed in Table 5.1. The parsers imported and used within this thesis are highlighted in gray. Also generic parsers for ontologies in the DAG-Edit and the OBO format are developed, where DAG Edit is the first attempt to save ontologies in a standardized format. Meanwhile it has been replaced by a format of the OBO (Open Biological Ontologies) initiative. As most important change the OBO format is not recursive anymore, which makes OBO flat files better human-readable. Both, the Gene Ontology (GO) and the Cell Ontology (CL) are in OBO format and imported by applying the OBO parser.

Step 2: Ontology alignment

Having external data sources imported into the ONDEX database, they all possess the same ontology based data format, i.e. all entities are stored as concepts linked by relations. For this reason all imported data sources are referenced from now on as ontologies, independent from the actual kind of imported data source.

This unified data format is exploited when the imported ontologies are aligned to each other. An alignment is a mapping of concepts from ontologies O_1 and O_2. If an alignment maps a concept c_1 of an ontology O_1 to a concept c_2 of an ontology O_2, then c_1 and c_2 are said to be *equivalent*. The mapping between two ontologies might be partial, i.e. O_1 and O_2 can obtain many non-equivalent concepts.

In ONDEX one concept of O_1 is mapped to another concept of O_2 according to the following three methods:

1. **Two-synonyms**:
   $c_1 \in O_1$ is equivalent to $c_2 \in O_2$, if they have at least two equivalent synonyms, i.e. if they have two or more of their names in common.

2. **Structalign**:
   $c_1 \in O_1$ is equivalent to $c_2 \in O_2$, if one of their ancestor nodes share the same name. This may be restricted to a certain depth within the ontology graphs and holds only for transitive relations.

3. **Transitive mapping**:
   $c_1 \in O_1$ is equivalent to $c_2 \in O_2$, if they are both mapped to the same concept of another ontology O_3, i.e. if $(c_1 \Leftrightarrow c_3) \wedge (c_2 \Leftrightarrow c_3)$, with $c_3 \in O_3$. Therefore, transitive mappings exploit the mappings generated by the previous two methods.

For all string comparisons the word stems of the concept names are used which are generated by a NLP tool in step 1.

Additionally, manual mapping lists (e.g. from the Gene Annotation Project (GOA) at `www.ebi.ac.uk/GOA`) can be used. An important feature of the alignment procedure in ONDEX is that here new links between equivalent concepts are generated rather then merging them into a new concept.

Further possible applications of the ontology alignment part of ONDEX are microarray data analysis and integration or gene annotation. For this purpose it is possible to use sequence analysis methods for the alignment of concepts that represent proteins and enzymes. However, sequence based mappings are not applied within this thesis.

5.1.3 Text mining

The text mining part of ONDEX is separated from the other parts and thus needs its own methods for text import, result generation and evaluation (Figure 5.1). However, the text mining section is connected with the other ONDEX modules in order to annotate the texts by concepts and to visualize text mining results using the ONDEX frontend (not applied within this thesis).

Although ONDEX can be used with any kind of text, in biomedical research often abstracts from the PubMed database MEDLINE are applied. These abstracts are freely available and contain a broad range of biomedical knowledge (regarding the selection of a text source see also Section 2.3.2). The MEDLINE abstracts are also used in this thesis and therefore we will refer in the following only to MEDLINE as text source.

Step 3: Import of texts

Importing texts requires mainly three sub-steps: at first it might be necessary to reduce the amount of texts by applying a pre-filter. Second, the text sources are parsed for the selected texts and tab-delimited text files for the actual database import are created. Finally, the texts are imported into the database and indexed by a full text indexing tool provided by the database.

Pre-Filtering: whether it is necessary to apply a pre-filter depends on the amount of texts to be mined and the available hardware resources. The MEDLINE 2005 database contains about 15 million abstracts, which is about 3.4 million more than in the preceding year. Considering all texts usually results in long database accessing times which make a reasonable use of the indexed text tables often impossible. Hence, a tool has been developed that processes the online PubMed query tool with a set of keywords and returns

the PubMed IDs (PMIDs) of the texts found. For this purpose, all synonyms and mapped concept names of the ONDEX ontology are used. The automated access of the Pubmed query tool with the concept names returns the same results as manual searches at the Pubmed search interface would do. The resulting HTML files are then parsed for the PMIDs of the abstracts.

Input file parsing and import file creation: following the filtering process, the MEDLINE files in the local repository are parsed for the selected texts and output files for the database import are created. MEDLINE files are delivered in XML format and contain additional information, as e.g. publication dates, author names, links to other databases and manually annotated MeSH terms (Table 5.1) characterizing the abstracts. Especially the MeSH terms are important since they are also imported as ontology into ONDEX and used for identifying the relations between concepts and texts. Thus, the MEDLINE XML files are parsed for the actual texts and useful additional information.

For each abstract parsed from the MEDLINE files (including the additional keywords) a copy is created that is processed by the same NLP tools used when databases and ontologies are parsed and imported (Section 5.1.2, step 1), i.e. each word is converted into its stem (*word stemming*) and added by a symbol indicating its most likely grammatical function (*part-of-speech tagging*).

Even if a filter is applied, the amount of text might still cause performance problems. Since large amounts of text generate large index files (created by the subsequently applied database text indexing tool), the time for searching especially common words in all texts can increase drastically. Therefore, in some cases not only one import file is generated, but rather this file is split into a number of smaller files. Correspondingly, in this case not only one `TEXT` table exists in the database, but as many as text import files are created. The information extraction algorithms accessing the texts are modified respectively, so that the existence of more than one table is detected and handled automatically.

Text import and database indexing: lastly, all import files are copied into the database and the text table is indexed using a full text indexing tool provided by the database (for details see Section 5.4 and the appendix, Section C). This enables a fast keyword search in all imported texts since applying regular expressions in the `LIKE` operator of the SQL `SELECT` command would not result in sufficient performance. The user can choose whether the original texts, the texts processed by the NLP tools or both are indexed.

Step 4: Concept based indexing

In this step, selected concepts are indexed in the texts, i.e. a list is created that maps each of the selected concepts to the texts in which it is found. It is important to note that this mapping is *concept based*: the ontology concepts including synonyms, different spellings and related concepts are identified rather than performing only string matches. The context of a concept can even be used to discriminate between different meanings of the same word. Of course, there is no approach resulting in 100 per cent accuracy and

there already exist many concurring algorithms. The concept based indexing approach described here can be characterized as dictionary based named entity recognition (NER) method (Section 2.3.3).

To index all imported concepts in all texts is only possible for small numbers of concepts and texts. In most applications, there will be at least two or three major ontologies imported (e.g. the MeSH terms, Gene Ontology and WordNet) and the MEDLINE abstracts. A concept selection is in most cases required for using the concept indexing. For this purpose, manually created lists containing the required ONDEX concept IDs can be used or any other method that selects concepts according to pre-defined criteria.

In the following it will be described how the names of this concept subset are identified in the texts (concept identification) and how the created mappings can be scored for discrimination between names of the same meaning (homonym detection).

Concept identification: for each selected concept all names are at first determined, i.e. all synonyms stored with the same concept ID and all names of all mapped concepts generate a list of terms to be searched in the texts. Duplicate names are not removed though they could be semantically different: a word sense disambiguation approach attempts to resolve this (see below).

The indexing algorithm takes each name from the list and uses the full text index created in the previous step 3 to retrieve IDs of all texts in which the word stems of the considered concept name appear. If the name consists of several words, only such texts are taken into further consideration where the words of the concept names appear consecutively in the same order.

A further criterion for rejecting a text is if only semantically different concept names are found, i.e. concept names that are not appearing on the search list. This might happen when these different concepts contain the name of the searched concept as substring. For example the concept "alpha amylase" appears as substring in the semantically different concept "alpha amylase activity". If "alpha amylase" is found in a text that solely deals with "alpha amylase activity", this text should be rejected. In order to prevent such wrong mappings, all concept names are checked for substring appearances in all other concept names.

Homonym detection: for each mapping a score for homonym detection is calculated and stored. Homonyms (in Greek homoios = identical and onoma = name) are words that have the same phonetic form (homophones) or orthographic form (homographs) but unrelated meaning. Thus, here we refer to homonyms as words or terms with different meanings, as e.g. "mouse" can be the animal or the computer device, and homonym detection can be seen as special task of *word sense disambiguation* (Section 2.3.3).

The basic idea to detect the correct semantics of a concept is to compare the ontology context of the considered concept with the context of the identified term in the text. The ontology context of a concept is the set of all names linked to the considered concept up and below in the ontology hierarchy until a certain depth. In contrast, the context of a

text is the set of all words appearing in the text (since all texts we apply here are only abstracts and rather short).

Then the resulting score is calculated as the ratio of the number of the words of the considered concept context contained in the text compared to the size of the text context. A score of 1 would mean that all words of the concept context appear in the text. The assumption is that the closer the score is to 1, the more the text deals with content related to the considered concept and hence, the more likely it is, that the text uses the considered concept in this specific sense.

For several reasons it turned out that this rule of thumb does not lead to robust results. The score depends strongly on the size of the texts and the chosen context depth. Some abstracts consist of only one or two sentences and thus are too short to contain enough context words for sufficient discrimination. The context depth on the other hand is also critical and depends on the size and the granularity of the ontology as well as on the position of the considered concept in the ontology. For some ontologies a depth of 3 might be sufficient. Using the same depth in other ontologies one might end up for most concepts at or close to the root concept of the ontology. This would generate a context too general, matching too many different texts. Thus, the proposed score for homonym detection deserves further improvements and is not further applied.

Step 5: Information extraction

The information extraction methods presented here are designed and developed for extracting intercellular signaling networks in MEDLINE abstracts, but have also been used in the diploma thesis of Rainer Winnenburg to extend a database of interactions between fungal pathogens and their hosts (Winnenburg, 2005; Winnenburg *et al.*, 2006). Generic implementations of these methods will be integrated into the re-designed ONDEX system currently under development.

The techniques applied to extract information from MEDLINE abstracts are *concept based co-occurrence searches* and *hypotheses generation*. As the concept based indexing described in the previous step, also these information extraction approaches are concept based and hence differ from usual approaches applied on a simple list of search terms (Section 2.3.4).

In the following we explain the concept based co-occurrence search and the hypotheses generation more detailed.

Concept based co-occurrence searches: In a usual co-occurrence search a list of search terms is applied. A *relation* between these terms is assumed if they appear concurrently within the same text (Section 2.3.4). For example, if two gene names "gene1" and "gene2" occur and additionally a keyword like "regulate" then a regulatory relation "gene1" → "gene2" (or "gene2" → "gene1", depending on the order of appearance of the keywords in the text) between the two genes could be assumed. In contrast, *concept based* co-occurrence searches assume a relation if two *concepts* are identified in the same text or sentence (using the concept based indexing, step 4). This has the following advantages:

- It is not necessary to define exhaustive word lists including all synonyms since a concept provides all equivalent names.
- Equivalent relations (i.e. relations between equivalent terms) are automatically collapsed into one relation since a concept name and its synonyms are not regarded as different entities.
- The known relations between the concepts can be exploited for mining new relations.

In order to perform a concept based co-occurrence search, a subset CS of concepts to be searched is selected and divided into different *concept groups* CG. In a concept based co-occurrence search then all possible concept combinations from a number of concept groups are created (*co-occurrence group*, COC) and checked for concurrent appearance in the texts.

More specifically, consider an ONDEX database with n imported data sources O_i (in the following referred to as *ontologies*), where each ontology consists of $|O_i|$ concepts. Then the set C of all concepts in ONDEX is defined as:

$$C := \{c_i^j \in O_i | 1 \leq i \leq n \text{ and } 1 \leq j \leq |O_i|\}, \tag{5.5}$$

where a concept c_i^j is defined as in Definition 5.2.

A concept subset $CS \subseteq C$ contains all concepts to be considered in the information extraction process. Such a concept subset is usually created by applying manually defined concept lists. Note that a concept $c \in CS$ consists additionally of the names from all concepts that have been aligned in step 2. Hence, if in the following co-occurrences are searched for two concepts $c_1 \in CS$ and $c_2 \in CS$ then this includes the search for co-occurrences of all concepts from C that possess an equivalence relation to c_1 or c_2.

Each concept $c \in CS$ possesses a *concept type* from the concept type set CT:

$$CT := \{t_1, t_2, ...\}. \tag{5.6}$$

The function ct then returns for each concept its type:

$$ct : C \rightarrow CT. \tag{5.7}$$

Using the concept types, the concept subset CS can be divided into *concept groups* CG_g, where each concept groups contains all concepts obeying a concept type $t_g \in CT$:

$$CG_g := \{c_s \in CS | ct(c_s) = t_g, t_g \in CT, 1 \leq s \leq |CS|\}. \tag{5.8}$$

To illustrate this, consider the cell-cell relation mining task (for details see Sections 5.2 and 5.3 below): concept types in this case are `cell` (cell types), `msngr` (messenger substances) and `rec` (receptor molecules), i.e. $CT :=$ {`cell`,`msngr`,`rec`}. The respective concepts are selected from all concepts in C (Definition 5.5) into a concept subset CS

and assigned to the correct type by using manually created lists. Finally, according to Definition 5.8 the concept groups contain all concepts belonging to a specific concept type. Thus, in this example a concept group CG_1 could be defined that consists of all cell type concepts $c \in CS$ with $ct(c) = \texttt{cell}$ and accordingly for the other concept types.

Subsequently, *co-occurrence searches* can be performed on combinations of concepts from different concept groups CG_g. Therefore, a *co-occurrence group* COC is a set of all concept tuples for which co-occurrences in texts are searched. Hence, a co-occurrence group is defined as the product set of a number of G selected concept groups CG_g:

$$COC := \prod_{g=1}^{G} CG_g = CG_1 \times ... \times CG_G = \{(c_1, ..., c_G) | c_g \in CG_g \text{ and } 1 \leq g \leq G\}, \quad (5.9)$$

i.e. the set COC consists of all tuples with ordered concept combinations from G different concepts groups CG_g, where G is at maximum the number of all concept groups. Such product sets contain no duplicate tuples. For example, in case of two concept groups $COC := CG_1 \times CG_2$ with each containing one concept $c_1 \in CG_1$ and $c_2 \in CG_2$, only the co-occurrences of (c_1, c_2) are searched and not the co-occurrences for (c_2, c_1).

To continue the cell-cell relation mining example, consider the search for co-occurrences of cell types and messenger substances in order to infer which cell types are able to release which ligands. Then the concept groups CG_1 and CG_2 can be chosen and the according co-occurrence group is $COC_{\text{cell-msngr}} := CG_1 \times CG_2$. Hence, $COC_{\text{cell-msngr}}$ contains all ordered pairs of concepts possessing the types `cell` and `msngr`. More specifically, for two exemplary cell types `cell` := {erythrocyte, hepatocyte} and one messenger substance `msngr` := {insulin} the corresponding co-occurrence group consists of the tuples $COC_{\text{cell-msngr}}$ = {(erythrocyte, insulin), (hepatocyte, insulin)}. A search for this co-occurrences group will return all texts mapped to both concepts of the first tuple (erythrocyte, insulin) as well as to both of the second tuple (hepatocyte, insulin) respectively.

Hypotheses generation: based on extractions of explicit knowledge stated in single texts, *hypotheses generation* as a second information extraction technique has the potential to generate new knowledge as well as to reproduce known facts by linking all relations sharing the same concepts. Relations occur on different levels, as e.g. on the level of explicitly described relations in a single text or as relation inferred from different texts. In the context of this thesis, *relation mining* is used to extract both kind of relations (see also Section 2.3.4 for a general introduction). First, concept based co-occurrence searches are used to infer relations described in single texts and subsequently, relations of concepts in different texts are reconstructed by using hypotheses generation (since complete cell-cell signals are usually not discussed in single texts). Thus, with hypotheses generation the *implicit* relationships between concepts can be discovered (see also Section 2.3.4).

Therefore, consider another co-occurrence search for messenger substances and receptors with $COC_{\text{msngr-rec}} := CG_2 \times CG_3$ and `rec` := {insulin receptor, IL-3 receptor}, resulting in co-occurrence searches for the tuples $COC_{\text{msngr-rec}}$ = {(insulin, insulin receptor),

(insulin, IL-3 receptor)}. If then, for example, co-occurrences are detected for both elements in $COC_{\text{cell-msngr}}$ and for (insulin, IL-3 receptor) $\in COC_{msngr-rec}$ these results are finally concatenated into two hypotheses: H_1 := (erythrocyte, insulin, insulin receptor) and H_2 := (hepatocyte, insulin, insulin receptor), i.e. "insulin" is the link to combine the co-occurrence results into a hypothesis and thus, relations between both cell types and the insulin receptor are presumed. Adding a third concept co-occurrence search for receptors existing in cell types will complete a cell-cell relation.

So the general approach is: all concept tuples of two co-occurrence groups CG_g and CG_h that could each be found concurrently in at least one text are combined into a hypotheses if they share a concept. The number of such concatenations between different co-occurrence searches is principally unlimited. That might lead in practice to a combinatorial explosion in the number of resulting hypotheses even for only small numbers of located co-occurrences. Hence, validation of the co-occurrence results and additional filters help to reduce the amount of hypotheses.

5.1.4 Graph analysis

Biological data is best seen as a set of interacting networks. This is also reflected by the basic data structure of ONDEX where concepts are linked by relations and thus, form a variety of graphs depending on the user's data selection. The ONDEX frontend is the interface to display and analyze networks from the data integration part as well as those extracted with the text mining module of ONDEX (Köhler *et al.*, 2006).

The ONDEX frontend (Figure 5.1) did already support the interpretation of gene expression results (Köhler *et al.*, 2006) and has also been used for visualizing and comparing bacterial metabolic networks. The development of the ONDEX frontend started when this thesis was nearly finished. So, the interaction between the ONDEX frontend and network extraction from text is not yet finished and will be included in the re-designed version of ONDEX.

Graph analysis and visualization with the frontend component of ONDEX works on an internal graph object which may be connected to arbitrary graph libraries as well as layout and filter algorithms by means of several interfaces and adapters. With this architecture a graph is generated from data imported from the ONDEX backend and subsequently passed to some algorithms independently of the origin of the graph. The results produced by these algorithms are transferred back into the internal graph object which then may be processed again by the available filter and layout algorithms. This way, arbitrary graph analysis and visualization processes are supported in order to provide the user with a wide range of possibilities for his specific application scenario.

5.2 Applying ONDEX to cell-cell relation mining

This section describes the application of the ONDEX workflow steps 1 to 5 (Figure 5.1) in the cell-cell relation mining project. Prior to the execution of this workflow, it has to be

defined for which entities relations are searched in the texts (Section 5.2.1). Following that, the import and alignment of relevant ontologies and databases (Section 5.2.2), the import of the texts to be mined (Section 5.2.3) and the indexing of the texts by the selected subset of the imported ontology concepts (Section 5.2.4) are performed. These are the necessary steps in order to execute the actual information extraction (Section 5.2.5).

5.2.1 Lists of searched concepts

As a basis for all subsequent processes, several lists that contain the concepts of interest are manually created by biomedical experts. The most important are: the cell type list (`cell`), the messenger substance list (`msngr`) and the receptor list (`rec`). These lists are composed of concept IDs from the Cell Ontology (CL) and the MeSH term (MESHD) ontologies (Table 5.1). Additional lists are taken from WordNet (WN) and contain concepts that describe the release of messenger substances from cells (`rword`), the binding of messengers to receptors (`bword`) and whether receptors are contained in cells (`cword`).

All concepts from these lists form the concept subset CS. The word lists as well as the entity lists can be found in the appendix (Section D, Tables 7.5 to 7.10). In the following presentation of all conducted steps, the symbols `cell`, `msngr`, `rec` as well as `rword`, `bword`, `cword` are used to denote the sets containing the entities of interest and important keywords respectively.

5.2.2 Import and alignment of ontologies and databases

Here step 1 (data source import) and step 2 (ontology alignment) of the ONDEX workflow (Figure 5.1 and Section 5.1.2) are executed. The ontologies and databases selected for import and alignment in the cell-cell relation mining project are (see also Table 5.1): the Cell Ontology (CL), the descriptors (i.e. the main headings) from the MEDLINE Subject Headings (MESHD), the Gene Ontology (GO), the Enzyme Nomenclature (EC), NCBI's taxonomy (TX) and the English language lexicon WordNet (WN). Furthermore, Biobase Transpath (TP) is imported and used for automated evaluation of the extracted ligand-receptor interactions. The import parsers for CL, GO, MESHD, WN are developed within the cell-cell relation mining project, including generic parsers for the ontology formats DAG-Edit and OBO which are also usable for any other ontology possessing these formats. The parsers for TX and TP were implemented by other members of the ONDEX project team and are applied here.

Problems during the data source import and the ontology alignment occur mainly at three different levels: for *syntactic*, *structural* and *special* conventions of the data sources.

Most common are *syntactical errors* in the data source. These could be for example inconsistencies in relations caused by simple typos which become apparent when both data sources are imported in ONDEX, e.g. if the ID of a GO term referenced by a MeSH term does not exist in GO. For manually annotated ontologies and databases any other type of misspelling might occur. In some cases, the flat files of some data sources did not obey their own format.

On a *structural level* more subtle import problems arise when a data source is interpreted in a way that was not originally intended. For instance the EC nomenclature was not designed as an ontology, but can be used like that if the hierarchical structure reflected in the ID numbers is exploited. For example, the enzyme class `3.4.11.6 ARGINYL AMINOPEPTIDASE` can be interpreted as a concept with an *is-a* relation to its parent class `3.4.11 Aminopeptidases` (and so on until the highest class level is reached). The problem at this point is that in some cases no parent classes are defined (which turned out to be an error occurring especially in the Swiss-Prot distribution of the EC nomenclature and was resolved after reporting).

Even more intricate are *special conventions* employed by the data sources. The Gene Ontology for instance applies the generic OBO format, but taxonomic specifications are defined within the names of the concepts. For example `GO:0007097` is "nuclear migration" and the concept `GO:0030473` is "nuclear migration (sensu Fungi)". The keyword "sensu" and the taxonomy term "Fungi" are given in parentheses and thus, a specification regarding an organism is included in the name of a concept. Embedding this information in the concept name complicates especially the automated alignment of ontologies since concept names are checked whether they denote the same entity. In our case we decided to remove the "sensu"-part from the concept name and to link the concept to a taxonomic entry in order to indicate the sensu relation. Therefore, corresponding taxonomic terms from TX (NCBI taxonomy database) are searched and an additional link between the concepts is created. Unfortunately, GO uses in the "sensu" phrase only the explicit organism names and not an ID number. Hence, misspellings in the names cause mismatches and can only be resolved manually. At the end, all terms including a sensu string are linked to such a taxonomy entry and the sensu information inside the concept name are removed.

Further problems occur when ontologies define finely granulated relation types that represent only slightly different contexts, which is for example the case in WordNet. Thus, despite the simple structure of the ONDEX database scheme, the problems differ depending on the data source and can in most cases only be solved by the individual parsers.

Finally, manually defined mappings between the concepts in the lists `cell`, `msngr` and `rec` defined in Section 5.2.1 are created. This is in some cases necessary to ensure mappings that have not been detected by the alignment algorithms. For this purpose, the alignment results for `cell`, `msngr` and `rec` concepts respectively are evaluated manually and complemented. The numbers of imported concepts per concept type and the resulting concept numbers (i.e. all equivalent concepts are aligned) can be found in the second and third column of Table 5.3.

Resulting from the various problems to be solved in the import process, a specific order has to be obtained for parsing the data sources, writing the import files and copying them into the ONDEX database.

5.2.3 Import of MEDLINE texts

In step 3 of the workflow (Figure 5.1 and Section 5.1.3) the texts to be mined are imported into the ONDEX database. Therefore, the MEDLINE files provided by the NLM with a

	Number of different texts	
(1)	MEDLINE 2005	14 792 864
(2)	After pre-filtering	3 484 760
(3)	Indexed with all subset concepts	2 875 284
(4)	Indexed with `cell`, `mnsgr` or `rec` concepts	2 335 656

Table 5.2: Number of all texts downloaded (1) and after pre-filtering with the Pubmed query tool (2). Applying then concept based indexing results in about 2.8 million different texts in which concepts from the manually selected concept subset are found (3) and about 2.3 million texts with identified concepts of biomedical texts (4).

total amount of about 50GB (for MEDLINE 2005) are applied. They are stored in XML format and contain not only the abstracts, but also additional information as e.g. pointers to other databases and manually annotated MeSH terms.

It turned out that indexing all MEDLINE abstracts (about 15 millions in MEDLINE 2005, see Table 5.2) is not feasible in reasonable time, depending on the number of concepts to be indexed and the available hardware resources. The most limiting factor seems to be the I/O performance, whereas CPU and RAM are not fully loaded. Thus, the full text indexing algorithm and the hard disk accessing times are crucial. For future ONDEX versions further full text indexing software will be tested, so that indexing of larger amounts of text might be possible.

Following that, at first the texts necessary to be imported are selected by applying a pre-filter. This is done by executing the Pubmed online query tool with the names of all concepts and their synonyms defined by the input lists `cell`, `msngr` and `rec` (Section 5.2.1).

Although the number of texts to be searched could thus be reduced to roughly 3.4 million by pre-filtering (Table 5.2), this turned out to be still too many for the database full text indexing tool if all texts are imported into only one database table. Terms like "activate" or "produce" for instance are usually used very frequently in biomedical texts. Especially in this pre-filtered subset of MEDLINE texts they occur nearly in each text. Hence, the index file generated by the database text indexer is very large and search queries based on this index are still very slow. As a consequence, the 3.4 million texts are further divided into 34 equally sized tables of 100.000 texts each and one table containing the remaining texts.

Technically, the steps 1 to 3 of the workflow considered so far are not executed strictly consecutive. Usually at first the import files for ontologies, databases and texts are created, followed by the actual import, ontology alignment and full text indexing. On the dual processor server used here these steps take approximately 30 hours, whereas the ontology alignment process, depending on the number of concepts to be aligned is usually the most intensive part (see Table 5.1 for imported concept and relation numbers).

5.2.4 Concept based indexing

The mapping of selected concepts to texts where the concept names appear, is the task performed in step 4 of the workflow (Figure 5.1 and Section 5.1.3). Therefore, the set of concept types used here is defined according to the Definition 5.6 and to the concept input lists (Section 5.2.1) as

$$CT := \{\texttt{cell}, \texttt{msngr}, \texttt{rec}, \texttt{rword}, \texttt{bword}, \texttt{cword}\}. \tag{5.10}$$

At first, the concepts possessing these types are iterated in order to get all belonging concept names. This includes synonyms and the names/synonyms of all aligned concepts. Furthermore, at this point only distinct word stems are considered since these are also accessed by the database text indexer. This helps especially to find concepts independent from their singular or plural form.

Next, a filter list is applied which contains concept names that are known to be inappropriate. Such mappings are not necessarily wrong and originate not only from false automatic alignments between different data sources, but also from single databases or ontologies. Especially the MeSH terms contain synonyms that would lead in our case to wrong indexing results. For example the term "Sheep alpha-Endorphin" is in MeSH a synonym to "alpha-Endorphin". Such terms are removed from the list of concept names to be searched in the texts.

Finally, the texts in the ONDEX database are searched for the remaining concept names. At this point all necessary information is compiled to perform the information extraction methods, i.e. the concept based co-occurrence searches and the hypotheses generation (step 5). Results gained in step 5 can be used as filter to reduce the number of texts, such that subsequent extraction tasks could be processed faster. Therefore, all steps (1 to 4) conducted so far can be repeated on smaller or different subsets of texts or concepts.

5.2.5 Information extraction

Based on the texts indexed by the entities of interest (cell types, messenger substances and receptors) the actual cell-cell relations can be extracted by performing concept based co-occurrence searches and subsequent hypotheses generation (step 5). Probably the most simple approach to begin with, is to search for texts containing concepts of all elements sufficient to describe a signaling relation, i.e. two cell types (source and target cell) and a first messenger. However, only very few texts deal with complete cellular interactions, as e.g. like "cell type A interacts with cell type B through messenger substance C". Even constraining the search by including keywords indicating interactions or signaling, as e.g. "interact", "release" or "signal", did not result in more specific texts.

Consequently, the co-occurrence search approach had to be refined. Reconsidering the biological background (Section 2.1), any cell-cell signal can be decomposed into three components (Section 3.1): the messenger release (source cell → messenger), the ligand-receptor binding (messenger → receptor) and the occurrence of receptors in cells

(receptor → target cell). Thus, each of these components can be searched independently by separate co-occurrence searches and subsequently combined into cell-cell relation hypotheses.

Double co-occurrence searches

After decomposing a cell-cell signal into its three components, the most straightforward approach is to search for double co-occurrences, i.e. tuples with each two concepts from different concept groups, in the abstracts. Therefore, according to Definition 5.8 three concept groups are defined by applying the previously defined set of concept types CT (Definition 5.10):

$$\begin{aligned} CG_1 &:= \{c \in CS | ct(c) = \texttt{cell}\}, \\ CG_2 &:= \{c \in CS | ct(c) = \texttt{msngr}\}, \\ CG_3 &:= \{c \in CS | ct(c) = \texttt{rec}\}, \end{aligned} \tag{5.11}$$

where the concepts c are from the concept subset $CS \subseteq C$, that is generated from the manually created concept lists (Section 5.2.1) and ct is the function that returns the concept type of a concept c (Definition 5.7). Thus, each concept group CG_g consists of concepts possessing a specified concept type $t_g \in CT$.

Using these concept groups, the co-occurrence groups containing all concept tuples to be searched are (according to Definition 5.9):

$$\begin{aligned} COC_{\texttt{cell-msngr}} &:= CG_1 \times CG_2, \\ COC_{\texttt{msngr-rec}} &:= CG_2 \times CG_3, \\ COC_{\texttt{rec-cell}} &:= CG_3 \times CG_1, \end{aligned} \tag{5.12}$$

Hence, each co-occurrence group contains a set of concept tuples, with each concept tuple consisting of two concepts from different concept groups. For instance, tuples in the co-occurrence group for `cell` and `msngr` could look like

$$COC_{\texttt{cell-msngr}} := \{(\texttt{erythrocyte}, \texttt{insulin}),\ (\texttt{hepatocyte}, \texttt{FSH}), ...\}.$$

Hypotheses are then generated by combining the tuples of the three co-occurrence searches possessing equal concepts connecting them, i.e. obeying the same `msngr` and the same `rec` concept.

Triple co-occurrence searches

To further restrain the co-occurrence search, three additional lists are applied which contain keyword concepts indicating that a text expresses the searched fact. These lists consist of concepts describing the cellular release or production of molecules (`rword`), the binding or interaction of molecules (`bword`) and when cells are able to contain or express molecules

(cword) respectively (Section 5.2.1). The respective concept groups (according to Definition 5.8) are:

$$\begin{aligned} CG_4 &:= \{c \in CS | ct(c) = \texttt{rword}\}, \\ CG_5 &:= \{c \in CS | ct(c) = \texttt{bword}\}, \\ CG_6 &:= \{c \in CS | ct(c) = \texttt{cword}\}. \end{aligned} \tag{5.13}$$

In contrast to the double co-occurrence searches, it is not of importance here *which* concept of a keyword concept group (Definition 5.13) is contained in a text, but rather if a text contains *any* keyword concept. For example, if a text containing the concept tuple (erythrocyte, insulin) is found, it is checked now whether additionally any keyword like "release" or "secrete" or "production" appears. Thus, it is not necessary to check for all concept triples that would be generated for e.g. $CG_1 \times CG_2 \times CG_4$, but rather for triples that have the set of all keyword concepts as third element. Therefore, we define new concept groups that contain each of the concept groups defined in Definition 5.13 as the *only* element:

$$\begin{aligned} CG'_4 &:= \{CG_4\}, \\ CG'_5 &:= \{CG_5\}, \\ CG'_6 &:= \{CG_6\}. \end{aligned} \tag{5.14}$$

Using these concept groups, the new co-occurrence groups for triple co-occurrence searches are (according to Definition 5.9) defined as:

$$\begin{aligned} COC_{\texttt{cell-msngr-rword}} &:= CG_1 \times CG_2 \times CG'_4, \\ COC_{\texttt{msngr-rec-bword}} &:= CG_2 \times CG_3 \times CG'_5, \\ COC_{\texttt{rec-cell-cword}} &:= CG_3 \times CG_1 \times CG'_6, \end{aligned} \tag{5.15}$$

i.e. each co-occurrence group contains now a set of concept triples, with each triple consisting as before (Definition 5.12) of two cell-, msngr- or rec-concepts and additionally a set of keyword concepts. So could, for example, a part of the co-occurrence group for cell, msngr and rword look like

$$\begin{aligned} COC_{\texttt{cell-msngr-rword}} := \{&(\texttt{erythrocyte}, \texttt{insulin}, \{\texttt{release}, \texttt{secrete}, \texttt{produce}, ...\}), \\ &(\texttt{hepatocyte}, \texttt{FSH}, \{\texttt{release}, \texttt{secrete}, \texttt{produce}, ...\}), ...\}. \end{aligned}$$

Thus, the texts found in the previous step are here checked for additional keywords. Hypotheses are then generated from the resulting reduced set of concept co-occurrences as before.

Triple co-occurrence searches in sentences

To get more specific results and to increase the precision rates, the texts gained by the two preceding steps (double and triple co-occurrence searches) are split into their single

sentences and searched again for triple co-occurrences. Technically, the same processes are applied as for the co-occurrence searches in whole texts: the sentences are regarded as "texts" (i.e., each sentence generates a single entry in the `TEXT` table (see also Figure 5.2) and indexed with all concepts of the concept subset. Finally, the searches for the concept triples of the previously defined co-occurrence groups (Definition 5.15) are performed.

Validation

The concept co-occurrences in each of the three tasks described above are validated by manual inspection of 100 randomly selected co-occurrence hits with one respective text each. The resulting precision value is the ratio of texts out of all sample texts that indeed describe the fact assumed by the co-occurrence. Although these might be too few evaluation samples, the precision values help to indicate tendencies.

For instance, a text containing a `cell` and a `msngr` concept is considered as correct if the text describes that this cell type is able to release this messenger. Any further conditions under which this signaling might take place are neglected. Thus, the "semantic range" of probably relevant texts is larger as if only special types of interactions are searched. That means that in case of messenger release not only texts describing a "release" or "secretion" of ligands are taken into account, but also texts talking about the "production", "synthesis" or "expression" of messenger substances. For our purpose it is assumed that cells which are able to produce a substance stated in the input lists as messenger substance are probably also able to secrete this substance. Similar assumptions hold for the other components as well: for ligand-receptor bindings texts also describing any "interaction" between both substances are positively evaluated and receptor occurrence in a cell can be characterized by the "expression" of the receptor or simply that a cell type "contains" receptor molecules. These assumptions are also reflected by the choice of additional keywords (see the lists for `rword`, `bword` and `cword` in the appendix, Section D).

Furthermore, a co-occurrence hit is rated as false-positive if one of the searched concepts does not occur at all, i.e. the concept-text match generated by the concept based indexing is incorrect, which is here not checked separately.

The validation measure used here is the *precision*, i.e. the proportion of extracted relevant entities to all entities retrieved (as defined in Section 2.3.1). Unfortunately, a recall value (i.e. the fraction of correctly identified entities in the set of *relevant* and thus true-positive entities) can not be measured since it is not known *a priori*, whether a text contains relevant information (regarding recall and precision measures see also Section 2.3.1). Note also that the same co-occurrence tuple can be selected for evaluation several times with different texts.

The generated hypotheses are difficult to evaluate for the same reasons as discussed in the database reconstruction approach (Section 4.5), i.e. many of them might hold true, but have not been investigated and reported explicitly yet. Hypotheses evaluation is best feasible for a small subset of cell types selected for a specific application (Section 6).

Concept type	Concepts	Names	Texts	Precision
cell	251 (283)	930	1 249 188	99.7%
msngr	196 (200)	1 765	1 285 035	99.8%
rec	178 (181)	1 502	362 722	99.1%
rword	11 (11)	21	974 562	81.2% (95.5%)
bword	7 (7)	28	1 091 127	94.5% (95.9%)
cword	8 (8)	21	921 019	88.4% (98.2%)
total	651 (690)	4 267	2 875 284	–

Table 5.3: Summary of the concept based indexing step of the cell-cell relation mining process with ONDEX (Section 5.3.1). For each concept type (column 1) the numbers of different concepts (i.e. including all equivalent concepts) identified in the texts (2) and defined generally (parentheses in (2)), the numbers of distinct concept names (i.e. word stems) belonging to these concepts (3), the numbers of different texts that contain concepts of the respective type (4) and a precision value (i.e. ratio of correctly identified texts) of an evaluated subset of each 1000 randomly selected texts (5) is given. Note that the same texts can contain several concepts and that therefore the total number of texts in the last line is not the sum of the numbers above. The two values for the different sets of words are explained in the text.

5.3 Results and validation

In this section the results of the concept based indexing 5.3.1 and the cell-cell relation mining performed using the indexed concepts (Sections 5.3.2 to 5.3.4) are presented. The relation mining is conducted in three subsequent tasks by applying the results gained in the previous task.

5.3.1 Text indexing results

Table 5.3 summarizes the results in this step for each concept type. For the number of different identified texts it can be observed that it differs largely from the number of pre-filtered texts (Table 5.2). Especially if texts containing `cell`, `msngr` or `rec` concepts are considered, the difference is about one million. This is surprising since the MEDLINE texts are pre-filtered using the same names of these groups (Section 5.2.3). Thus, this could mean that the ONDEX concept based indexing could identify a qualitatively better set of texts or that the approach misses texts. To elucidate this question, for each concept group a randomly selected set of 1000 concept/text pairs are evaluated semi-automatically. First, all selected texts are checked automatically whether they contain exactly one of the names of the assigned concept. As second step, all remaining texts are validated manually.

The resulting precision values are the ratios of all correctly identified concepts to the number of samples. These are quite high for the biomedical concepts (`cell`, `msngr` and `rec`), i.e. the absolute numbers of false positives are 3, 2 and 9 respectively. In case of the word lists (`rword`, `bword` and `cword`) the main problem is the discrimination between nouns, verbs and adjectives since only word stems are used and this differentiation might get lost in some cases (e.g. “secrete”/”secretion” or “produce”/”production”). But the intended

COC	Maximal possible hits
`cell-msngr`	49 196 (56 600)
`msngr-rec`	34 888 (36 200)
`rec-cell`	44 678 (51 223)

Table 5.4: Maximal possible hits for each co-occurrence group. These numbers are the products of the concept numbers in each respective concept group (see Table 5.3, column 2), whereas the first number is the product of all concepts that could be identified in the texts and the number in parentheses is the product of all defined concepts.

meaning is mostly the same and the texts are still sufficiently identified for our purposes. Thus, the precision values of the word concept types preceding the parentheses in Table 5.2 determine the "strong" evaluation, i.e. the part-of-speech recognition is considered, whereas the values in parentheses measure only if the word sense is met.

The complete concept based indexing process took on the double processor machine about 14 days for 4267 concept names to searched in 3.4 million texts.

5.3.2 Double co-occurrence searches

For all combinations of the 251 `cell`, 196 `msngr` and 178 `rec` concepts identified in the texts (Table 5.3) co-occurrences are searched in about 2.3 million texts. Table 5.4 shows the maximum numbers of possible co-occurrences and the first row of Table 5.5 shows the resulting numbers of co-occurrence hits, texts and hypotheses. In each co-occurrence group of the double co-occurrence search about a third (or less) of all possible co-occurrences are detected, the number of texts also decreases strongly compared to the 2.3 million texts in which single occurrences of the three according concept groups (Definition 5.11) are found. The total number of different texts (last column in Table 5.5) is about 22% of the original text volume.

Hit and text numbers are not as important as the actual precision with which the searched facts are identified. Regarding the double co-occurrences, the precisions are relatively low for the co-occurrence groups $COC_{\texttt{cell-msngr}}$ and $COC_{\texttt{msngr-rec}}$ and significantly higher for the receptor detection with $COC_{\texttt{rec-cell}}$ (Table 5.5). The number of 107 million hypotheses (Table 5.5, col. 6) is much less than the number of all concept combinations possible in principle, but, however, obviously too much for manual examination or any further use.

Two questions directly arising here are 1) if and how the quality of the results can be improved, especially for the first two co-occurrence groups and 2) whether the distribution of the precision values among the three co-occurrence groups is an arbitrary effect or might remain stable in different searches. Therefore, the co-occurrence results are attempted to be improved by adding further keywords in the next search task which would probably also affect the number and quality of the resulting hypotheses.

Text type	COC	Hits	Texts	Prec.	Hypotheses	Total texts
Abstract	cell-msngr	16 386	332 475	15%	107 135 827	515 555
	msngr-rec	11 420	191 648	9%		
	rec-cell	11 593	113 102	45%		
Abstract	cell-msngr-rword	13 910	170 202	15%	94 843 489	289 578
	msngr-rec-bword	9 372	97 243	1%		
	rec-cell-cword	10 061	64 637	41%		
Sentence	cell-msngr-rword	6 334	97 092	22%	3 898 341	126 916
	msngr-rec-bword	2 213	15 917	12%		
	rec-cell-cword	3 230	15 137	70%		

Table 5.5: Overview of the main text mining results for cell-cell relations after three steps (column 1: abstracts, unfiltered (first row), abstracts, filtered (second row) and single sentences (third row). Each relation mining step is divided into three independent co-occurrence searches for the components of a cell-cell signal (2). For each co-occurrence search, a number of co-occurring concepts (3) are found in a number of different texts (4). A precision is determined by manual examination of 100 randomly selected co-occurrence hits (5). The three components of a cell-cell signal are combined into complete cell-cell signaling hypotheses (6). The last column (7) is the total number of different texts resulting from all respective co-occurrence searches. Note that in the third section (Sentence) single sentences are regarded as the texts.

5.3.3 Triple co-occurrence searches

Compared to the double concept co-occurrence searches, the number of triple co-occurrence hits decreases as well as the number of texts in which these co-occurrences are found, whereas the text number reduction is more significant than the decrease in the number of co-occurrence hits (second row of Table 5.5). Also the number of resulting hypotheses is lower as for applying double co-occurrences. However, the reduction is only about 11%, which means that still too many hypotheses remain for reasonable manual investigation or further network analysis.

The precision values, obtained for 100 randomly selected samples with the same conditions (as described in Section 5.2.5) did not improve. In two cases they even lowered. What remains is the distribution pattern of the three precision values: the ligand-receptor binding search performs worst, the search for receptors in cell types best and the messenger release is roughly in the middle of both. Manual inspection of the texts revealed that most of the false-positives are due to the fact that many texts mention a number of substance and cell type names in very different contexts. Some abstracts are quite long and some contain long lists of names (e.g. with substances that have been investigated).

Thus, to draw from a co-occurrence hit the conclusion that the respective text describes the fact assumed would be in most cases highly speculative, even if additional indicating keywords are mentioned. However, what could also be seen is that many of the searched facts are described in single sentences of the considered abstracts and hence, the next step is to consider all sentences of the remaining texts individually.

5.3.4 Triple co-occurrence searches in sentences

First, 2 992 991 individual sentences from the 289 578 remaining texts (Table 5.5) are extracted and indexed by the concepts as in the previous tasks, resulting in 1 910 309 sentences where searched concepts are found. Using these sentences, the concept co-occurrences gained here (third row of Table 5.5) show in all categories a much higher difference to triple co-occurrences in abstracts than comparing triple with double co-occurrence searches. The number of co-occurrence hits is reduced between 25% to 50% of the triple hits in abstracts, depending on the co-occurrence group. The text numbers can not be compared, since here "texts" are single sentences. But considering that from about 2 million indexed sentences only some 120 thousands remain (about 6%), this is also a great reduction. Likewise noticeable is the reduction in the number of resulting hypotheses (only about 4% of the number of hypotheses from triple searches in abstracts).

Remarkable changes also occur in the precision values. All rates increased by sustaining the pattern emerged in the two previous co-occurrence searches. The values shown in Table 5.5 are based on 100 randomly chosen results, but for sentences additionally evaluations for 300 random samples are performed. Here, no significant differences can be observed compared to the 100 random samples (i.e., the values are 23%, 14% and 67% for the three co-occurrence groups respectively).

Thus, compared to the previous tasks it can be observed that the sentence-based co-occurrence searches are the most successful ones.

5.4 Implementation

The central component of the ONDEX system implementation is the object-relational database management system (DBMS) PostgreSQL. Running on a dual processor server with 2 Intel Xeon 2.8 GHz running under Linux and controlled by Makefiles and Shell Scripts, the database scheme shown in Figure 5.2 is created and serves as base for all data integration and information extraction processes. The programming languages SQL, Java and ANSI C are used to perform the ONDEX functions as data parsing and import, ontology alignment, concept based indexing and information extraction. Furthermore, additional tools for full text indexing (TSearch2), word stemming (Snowball) and part-of-speech tagging (QTag) are applied. All programs additionally developed for the cell-cell relation mining are implemented in Java, access the DBMS via JDBC and are controlled by Shell Scripts and Makefiles.

5.5 Discussion

The main characteristic of the ONDEX framework and the text mining approach applied here is that they are *concept based*. In concept based approaches the texts are annotated by *concepts* from other knowledge representations, as e.g. controlled vocabularies or ontologies (Tan, 1999). Using this background knowledge, information on the semantics of a

text is available and known relations of the concepts can be used to extract information from text.

In contrast to other concept-based approaches, ONDEX does not apply only one data source for annotating the texts (as e.g. in biomedical text mining often UMLS is used, see Vintar *et al.*, 2003; Hofmann and Schomburg, 2005), but an arbitrary number of ontologies and databases can be imported and aligned. These integrated ONDEX ontologies are used in all subsequent processes to support text mining at a semantic level.

A similar approach regarding text mining for ligand-receptor interactions is applied by Albert *et al.* (2003). In this work, a subset of MEDLINE texts is selected, tagged by terms from a dictionary (protein names, keywords) and finally triple occurrences at the sentence level are extracted. They use a hierarchically organized dictionary that consists of relations between the terms (similar to ontologies). The dictionary is manually created and subsequently extended by the results gained from the text mining. An important difference to our approach is that we do not rely on only one manually created ontology, but rather apply and map standard databases and ontologies automatically.

In the following, the main results gained from concept based indexing, concept based co-occurrence searches and hypotheses generation for cell-cell relations are discussed in more detail as well as possibilities for future work.

Concept based indexing

The concept based indexing is best characterized as dictionary based named entity recognition (NER) approach (Section 2.3.3) with the specialty that the applied dictionary is generated by integrating many different sources and thus provides a wider background knowledge than a single dictionary could. Also word-sense disambiguation could be feasible using the context of concepts in the ONDEX ontology. Initial tests of such a homonym detection were not successful here (Section 5.1.3, step 4), but the consideration of further conditions, as e.g. the length of the texts, will probably help to improve this approach. Hence, improving the homonym detection is one of the important next steps.

The quality of our concept based indexing NER approach is evaluated for a selection of random samples and shows high precision rates (Table 5.3). Even if the partly failed part-of-speech recognition is taken into account, the precision is still greater than 80%. So it can be assumed that the difference in the number of texts returned by the Pubmed pre-filter with the number of texts mapped to `cell`, `msngr` and `rec` concepts (Table 5.2) is not due to failed concept recognition in the texts. The selected texts contain the searched entities with a high probability and are thus a qualitatively good base for the subsequently applied text mining.

Concept based co-occurrence searches

One of the main results regarding the co-occurrence searches for concepts is the stability in the distribution of the precision values among the three different searches in all settings (Table 5.5). Although the number of the randomly selected samples is not exhaustive, a

clear tendency becomes visible: in both abstracts and sentences, the messenger-receptor interactions are difficult to identify whereas the detection of receptor expression in cells performs best. The messenger release from cells takes a position in between the other searches by performing significantly better than the messenger-receptor interaction search. Thus, the quality of the results seems to be dependent on the searched relation type. This is similar to results reported by Ding *et al.* (2002), who received better co-occurrence results on the sentence level for only some relation types.

Manual inspections of the search results reveal that the high precision value of $COC_{\texttt{rec-cell-cword}}$ (compared to the rates of the other two searches) is mainly due to the stability of the formulations in the sentences used to describe that a cell type contains a specific kind of receptor molecules. This is the case if the receptor is expressed in this cell type and this in turn is mostly stated as e.g. "X cells express receptor Y", "receptor Y is expressed by X cells" or "Y receptor expression on X cells is ..."(with X and Y as the names of the cell type or the receptor respectively). Furthermore, such sentences are often short and contain only few more molecule or cell type names, which is often the reason for false-positives. Here, a formulation convention seems to have emerged.

The ligand-receptor binding detection at the other side of the precision value spectrum suffers from the varying possibilities to describe the interaction of two molecules. For example, often the receptor name is not explicitly mentioned (e.g., "the binding of messenger X to its respective receptor") and is thus harder to detect. Messenger release from cells as the last remaining co-occurrence group, does not achieve as good rates as for receptor expressions, but perform significantly better then messenger-receptor interactions. For the messenger release it can be observed that the formulations used in the texts are relatively stable (e.g., "enhances the production of Y molecules in X cells"), but not always unambiguous, as e.g. in "introducing Y molecules enhances the production of X cells", where the direction of causality has changed. The application of further filters and the consideration of the relative position of the concepts in the sentences might help here (see below).

It is controversially discussed in the literature whether co-occurrence approaches might be too simplistic or whether they are generally able to produce reliable results (Chen and Sharp, 2004). Some co-occurrence approaches were indeed successful, especially in the context of extracting gene networks (Jenssen *et al.*, 2001). Our observations might help to decide whether to apply a co-occurrence search or not: since the advantage of co-occurrence approaches is their simplicity, i.e. they are easy to apply and perform faster than parsing each sentence for its grammatical structure, they could be applied as first attempt and evaluated. Examining even small sample sets is usually sufficient to get an intuitive feeling about the feasibility of a co-occurrence search in the respective case. Then it can be decided whether further simple co-occurrence filters would help or if a deeper natural language analysis should be performed. In case of the application of more intensive parsing techniques, co-occurrence searches are useful to reduce the amount of text.

Summarizing the co-occurrence result discussion, it can be concluded that the sentence-based co-occurrence searches are the most successful ones. The indexing of the concepts in the texts ensured that the searched entities appear in the selected text with high probability. Also detecting equivalent relations is made possible by combining all equivalent

concepts into one. Furthermore, it can be assumed that most of the missing hits from the triple co-occurrence searches in abstracts are probably false-positives since the precision values increased strongly when a sentence based search is applied. This confirms results from other work, where good performance on sentences compared to the text level is also reported (Ding *et al.*, 2002). Nevertheless, both co-occurrence runs conducted first are necessary to reduce the text amount and thus, computation time. Also they helped to gain an intuitive sense about the text structures and ideas for further reductions (see below).

Hypotheses generation

The concatenation of co-occurring concepts into complete cell-cell relation hypotheses leads (due to combinatorial explosion) naturally to large hypotheses numbers and is considered as serious problem for hypotheses generation approaches in general (see Weeber *et al.*, 2005, and also Section 2.3.4). In our case it is difficult to calculate a score for automated hypotheses ranking (as applied in hypotheses generation, see e.g. Srinivasan and Libbus, 2004; Wren *et al.*, 2004) since appropriate additional information is lacking. Thus, the strategy employed here is to remove presumably false-positive co-occurrence hits by sequential application of co-occurrence searches with increasing restrictions. This finally reduces the number of hypotheses to about 4% of the initial number (Table 5.5), but 4 million hypotheses are obviously still too many for manual consideration. Therefore, additional filter steps on the co-occurrence level could be applied in future work (see below).

The application of the generated hypotheses considered here is to use a part of the remaining cell-cell relation hypotheses in order to gain information about cellular communication in neurodegenerative diseases (Section 6). For this purpose, the cell types will be restricted to the four cell types of interest in the context of these diseases. The number of co-occurrences and resulting hypotheses are then sufficiently low to allow exhaustive evaluations.

Future work

Future work in extraction cell-cell interactions from text might comprise a variety of different methods. The methods proposed here consider mainly the improvement of the quality and number of concept co-occurrences since these are feasible to evaluate and a reduction of their number will reduce also the number of the hypotheses, which usually are difficult to validate. Finally, all hypotheses could be made accessible for search and inspection by a new database that incorporates the text mining results.

Co-occurrence frequencies: the frequency of co-occurrences, i.e. the number of different texts in which the same co-occurrence appears, was not accessed yet. The assumption here is that frequently appearing concept tuples can be expected to have a higher probability of being related than concepts appearing concurrently only once or a few times, since they are often used together in the same text. In the present context, it was problematic to apply co-occurrence frequencies due to the large total number of texts, which would lead

to relative low frequencies even for concepts co-occurring in several thousand texts. Also, most co-occurrences appear in less than thousand different texts. So it would be necessary to compare absolute hit numbers instead of frequencies or to use a different base for the calculation of relative frequencies. If such a measure could be implemented, it would help to rate the quality of co-occurrences, and thus also to score the hypotheses resulting from the co-occurrences. Selections of presumably high quality results could then be evaluated first.

Automated validation by external data: to further improve evaluation and subsequent hypotheses combination, existing information of molecular databases can be accessed. In ONDEX the molecular interaction database Transpath (Schacherer *et al.*, 2001) is available (Table 5.1) and can be exploited to evaluate the extracted messenger-receptor relations ($COC_{\texttt{msngr-rec-bword}}$), as this seems to be the most complicated co-occurrence search. Therefore, from the 9 372 `msngr-rec` triple co-occurrences identified in abstracts (Table 5.5), 47 could be located in Transpath. This low number results mainly from the fact that Transpath is not a very large database (Table 2.1 in Section 2.2), and not many ligand-receptor interactions are incorporated. But applying these valid interactions reduces the number of resulting hypotheses to only 397 748 (compared to about 94 millions of all extracted `msngr-rec` interactions). Working with these hypotheses for further examinations might leave out correct ones, but constrains the selection to only such hypotheses based on a valid ligand-receptor interaction.

Co-occurrence filters: a possibility to exclude false-positive co-occurrence results is to examine the texts for formulations that can be used as indicator for removing these texts. For example, in case of the search for messenger release relations with $COC_{\texttt{cell-msngr-rword}}$ in many sentences that contain the words "effects of" or "effects on" not the release of a `msngr` *from* a `cell` is reported, but rather the effects *of* the `msngr` *on* a `cell`. Thus, the causality indicated by `cell` → `msngr` is switched.

A filter that removes all sentences containing these phrases was tested and evaluated for 300 sample sentences selected at random. The precision rate did indeed increase a little to about 29% (compared to 22% in the triple co-occurrence search on sentences, Table 5.5). But evaluation of another 300 samples selected at random from the 22 235 *removed* results showed a precision of about 20%, meaning that about a fifth of the deleted results might be correct and are thus wrongly removed. One reason for this is that for the concept based indexing the WordNet concept `WN:13500435NN` with "effect" as the only concept name was used. This caused indexing of all sentences containing the words "express" or "expression", which probably affected too many. Also, in all approaches applied here, the order of the concepts in the text is not taken into account.

However, this does not discount filtering approaches in general, but shows that a possibility to survey quickly the resulting effects of a filter will help to remove false-positives.

Co-Occurrence templates: following from the previous suggestions, an important improvement would be the possibility to apply templates and simple extraction rules on the selected sentences. From the order of the searched context and their relative positions to relevant keywords, probably more accurate conclusions about the most likely content of the respective sentence could be inferred. The success of such rule-based approaches has also been shown in text extraction competitions (as discussed in Section 2.3.4 and in Yeh *et al.*, 2003). For this purpose, the concept based indexing in ONDEX should be complemented in a way that the positions of the indexed concepts in a text are also considered.

This would probably also help to avoid another problem that became visible through manual examinations of the false-positives: many sentences consists of long lists of substance, molecule, cell type or tissue names. For example in the sentence "A murine model was developed to assess the direct and indirect effects of murine IL-2 and the secondarily released cytokines, gamma interferon (INF gamma), and tumor necrosis factor (TNF alpha), on testosterone production in isolated Leydig cells." (second sentence in Meikle *et al.*, 1992), the co-occurrence of "testosterone" and "Leydig cells" is a true-positive messenger release, whereas any other combination of the cell type with the messenger substance names is false-positive in this context.

Such lists can be even longer: "In this study, we investigated the effects of IFN gamma on the production of cytokines (IL-6, IL-8, IL-10), prostaglandin E(2)(PGE(2)), proteoglycans (PG), nitric oxide (NO), interleukin-1 receptor antagonist (IL-1ra) and stromelysin by non-stimulated and IL-1 beta-treated human chondrocytes." (second sentence in Henrotin *et al.*, 2000). Obviously, not any pair of these names reflects the correct semantic of the sentence. But if the position of keywords like "produce" or "binds to" is known in relation to the list of names, it can be better approximated which entity affects which other entities.

Hypotheses database: a possible application of the hypotheses gained so far is the creation of a database that contains all extracted potential signals in order to be queried by biomedical experts searching for new ideas regarding specific questions on intercellular signaling. Such a database could be periodically updated by automated ONDEX processes that download recent MEDLINE texts, import and index them, and finally perform the co-occurrence searches and hypotheses generation. A search in the texts that remain after the application of the full text mining process would be much more specific than a simple query for the respective entity names at the Pubmed interface.

Chapter 6

Hypotheses generation for neurodegenerative diseases

Contents

In this chapter the cell-cell signaling networks extracted from the CSNDB database (Section 4) as well as from MEDLINE abstracts (Section 5) are investigated if they could be applied to more specific biological questions related to neurodegenerative diseases. Hence, instead of trying to reconstruct and analyze the whole intercellular communication network, only a small subset of cell types is considered here.

Therefore, an introduction to neurodegenerative diseases and the special focus of the research conducted by the group of Thomas Schmitt-John at Bielefeld University is presented in Section 6.1. This section also includes a description of the affected cell types as well as some phenotypic effects that cell-cell signals might cause in this context. In Section 6.2 then it is shown how the previous results are restricted, evaluated and how the subnetwork of the selected cell types finally is generated. The resulting hypotheses are partially evaluated and the results are discussed in Section 6.3.

6.1 Intercellular signaling in the context of neurodegenerative diseases

Neurodegenerative diseases are hereditary and sporadic conditions which affect the brain function and are characterized by a progressive nervous system dysfunction (Beal *et al.*, 2005). These disorders result from deterioration of neurons and are often associated with atrophy of the affected central or peripheral nervous system structures. They are divided into two groups:

1. conditions causing problems with movements
2. conditions affecting memory and conditions related to dementia

Such neurodegenerative diseases are e.g. Alzheimer's Disease, Parkinson's Disease, Multiple Sclerosis, Amyotrophic Lateral Sclerosis (ALS or Lou Gehrig's Disease) or Huntington's Disease. Paradoxically, neurodegeneration is a major element in many diseases that are often not usually classified as degenerative (e.g. multiple sclerosis or epilepsy) and conversely, inflammatory processes for example are activated and vascular compromise occurs in some degenerative diseases (Williams, 2002). Age is the major risk factor for neurodegenerative diseases and as society ages, neurodegenerative diseases will become increasingly common. Thus, research conducted in this field is becoming of increasing importance as well.

At Bielefeld University Thomas Schmitt-John's research group is researching the so-called *wobbler mouse* (Falconer, 1956), a mouse mutant serving as model organism for the human amyotrophic lateral sclerosis (ALS) disease. Together with two researchers of this group, Thomas Schmitt-John and Carsten Drepper, the cell-cell communication networks derived by reconstruction from databases (Section 4) and text (Section 5) are inspected whether they can be applied in the context of neurodegenerative diseases. The special focus of the research group is the wobbler mouse and the respective human disease which will both be briefly introduced in the following section (Section 6.1.1). However, the goal of the reconstruction application is to shed light on the communication between cell types affected in neurodegenerative diseases in general. Therefore, the respective cell types are presented in Section 6.1.2 and examples of typical phenotypic effects are shown where cell-cell signals probably play a major role.

6.1.1 The wobbler mouse and ALS

The wobbler mouse was first described by Falconer (1956) as a spontaneous mutation, that was characterized by their wobbly gait, smaller size and fine tremor of the head (Duchen and Strich, 1968). In the following years it became the most popular murine model of motoneuron diseases, characterized by the progressive degeneration of motoneurons, resulting in muscular weakness, paralysis and death.

The most common adult human motoneuron disease is amyotrophic lateral sclerosis (ALS). ALS is a rapidly progressive neurological disorder, characterized by a rapidly progressive degeneration and loss of motor neurons in the brain and spinal chord, which ulti-

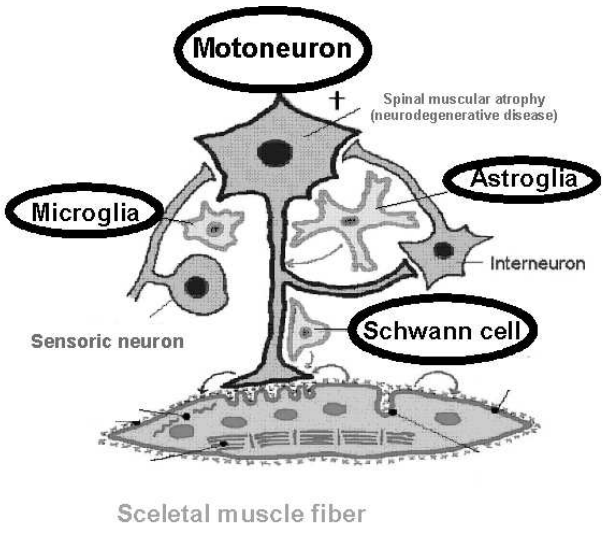

Figure 6.1: Cell types mainly affected in neurodegenerative diseases: motoneurons, microglia, astroglia and Schwann cells. In case of ALS and the wobbler mouse, the motoneurons that connect the brain with the sceletal muscle fiber are degenerating and finally die. The surrounding cells proliferate during the motoneuron degeneration (probably in order to support the dying cells). Figure with permission by Harald Jockusch (Bielefeld University).

mately leads to paralysis and premature death. The prevalence of ALS is approximately 5/100,000 in the population and increases with age (peak at age of 60-75: 33/100,000 for men, 14/100,000 for women, see Majoor-Krakauer *et al.*, 2003). Common clinical features of ALS include muscle weakness, fasciculations, brisk (or depressed) reflexes, and extensor plantar responses.

Most ALS cases are of unknown etiology, so, although the chromosomal localizations of the mutations found in the wobbler mouse show no similarity to the few human forms of ALS for which the mutation is known, the wobbler mouse remains the main source to explore this disease (Boillée *et al.*, 2003). The wobbler mouse is used to explore the course of neurodegeneration at the clinical, cellular and molecular levels.

Further aspects of the wobbler syndrome were discovered by Leestma and Sepsenwol (1980) and Heimann *et al.* (1991) who found that the neurodegenerative syndrome is associated with male sterility. Augustin *et al.* (1997) then explored the hypothesis of a humoral factor in the wobbler mutation and showed the cell acting autonomous of the wobbler mutation. Schmitt-John *et al.* (2005) identified the mutation in the Vps54 gene as the cause for motor neuron disease and defective spermiogenesis in the wobbler mouse.

6.1.2 Cell types affected in neurodegenerative diseases

Cell types mainly affected in neurodegenerative diseases are motoneurons, microglia, astroglia and Schwann cells (Figure 6.1). Communication between these four cell types plays an important role in different diseases. Although the cell signaling hypotheses extracted for this purpose support the research on several diseases, we will give here exemplary details

	All	Neuro	Neuro-Pos.	Prec.
`cell-msngr-rword`	6 334	241	114	47%
`rec-cell-cword`	3 230	195	140	71%
`msngr-rec-bword`	2 213	911	82	<10%
hypotheses	3 898 341	347	–	–

Table 6.1: Text mining results after restriction to cell types affected in neurodegenerative diseases. The rows show the numbers of hits found in each triple co-occurrence type and (in the last row) the number of hypotheses resulting from these co-occurrences: for all cell types (column 2, see also Table 5.5, Section 5.3), for the four cell types in neurodegenerative diseases (3), the number of positively evaluated results (4) and the respective precision (5). In this case, the co-occurrence hits and hypotheses are completely examined manually. Note, that the number of the third co-occurrence search, `msngr-rec-bword`, is restricted to all positively evaluated results of the other two co-occurrence searches.

of disease processes in which cell communication might play a crucial role that are related to the wobbler mouse and the human ALS disease.

In the center of these diseases are motoneurons which transfer motion control signals from the brain through the spinal cord to the muscles. Influences of intracellular signals have been investigated for instance by gene expression profiling of receptor molecules (Festoff *et al.*, 2000) or by measuring the abundance of substances in the cerebrospinal fluid (Brooks *et al.*, 1983).

A further possible influence of cell-cell signals could be related to the reactions of the cell types surrounding the motoneurons, as the astroglia and the microglia. These cells are observed in the wobbler mouse to proliferate when the motoneurons have been affected. The reason is probably the attempt to support the degenerating cells. The question, however, is how the surrounding cells "know" that the motoneurons are dying. It is an obvious suggestion that here cellular communication might be involved.

Thus, these few examples demonstrate that intercellular signals are probably part of the complex disease processes. Therefore, the results gained for all cell types will be restricted in the following to the four presented cell types in order to better understand their communication behavior and involvement in the diseases. Furthermore, another benefit of this restriction will be the possibility to test the quality of the applied methodology in a data set of manageable size.

6.2 Resulting cell-cell signaling hypotheses and evaluation

6.2.1 Applying the database results

Investigating the networks reconstructed from the CSNDB (Section 4) reveals that from the few cell types contained at all in CSNDB, the Schwann cells appear as the only of the four cell types of interest. The other cell types are mostly part of the immune system

(as e.g., natural killer cells, T helper cells etc.). Or the available location sites are too unspecific for our purposes (as e.g., nerve cell). This situation is similar for the HPRD (Sec 2.2). Although this recently published database seems to be more accurate and complete compared to the CSNDB, it lacks also most of the cell types of interest here. Thus, the previously extracted CSNDB network could not be further applied and additional effort to apply the HPRD would not result in applicable data for this specific context.

6.2.2 Applying the text mining results

In order to acquire hypotheses from the text mining results for only the four cell types affected in neurodegenerative diseases (according to Section 6.1.2), the following steps are performed:

1. The co-occurrence results for triple co-occurrences in sentences that contain cell type concepts (i.e., `cell-msngr-rword` and `rec-cell-cword`) are restricted to the four selected cell types. These restricted sets of sentences with co-occurrences are then completely evaluated.

2. The co-occurrence results for `msngr-rec-bword` are restricted to only those combinations of `msngr` and `rec` that occur in positively evaluated co-occurrence results from step 1. Also these `msngr-rec-bword` co-occurrences are completely evaluated.

3. Hypotheses for all positively evaluated co-occurrence results from step 1 and step 2 are finally generated and available for manual inspection.

To allow a complete evaluation of all co-occurrence results, the order of these steps is important: by accessing only those `msngr-rec-bword` results that are connected to positively evaluated `cell-msngr-rword` and `rec-cell-cword` sentences the number of `msngr-rec-bword` hits could be reduced from 2 213 to 911 (Table 6.1). Thus, using this reduction step, a complete evaluation of the `msngr-rec-bword` co-occurrence results became feasible.

The evaluation of the co-occurrence results by biomedical experts is conducted by comparing the conclusion drawn from a co-occurrence (e.g., that a cell type is capable of producing or secreting a specific first messenger) to the sentence where this co-occurrence is found. A co-occurrence hit is positively evaluated if the conclusion is described in the respective sentence. Since the triples of terms can occur concurrently in several sentences, a maximum of three randomly selected sentences are manually inspected. If in at least one case the conclusion is described by the sentence, the co-occurrence hit is evaluated as positive.

The resulting precision values of the concept based co-occurrences can be found in Table 6.1 For comparison, the previously gained values for all cell types (see Table 5.5 in Section 5.3) are listed in the second column. The respective positively evaluated co-occurrence hits are listed in the Tables 7.11 to 7.13 (appendix, Section E).

The number of finally resulting hypotheses is 347 from which 100 were manually validated. This validation resulted in about one-third true-positives. The remaining hypotheses are rated as false-positives for two main reasons:

1. The concepts often include too many synonyms from conceptually different levels, as e.g. substance names and names of substance classes. Such names should not be regarded as synonym. This is caused mainly by improperly defined MeSH terms and MEDLINE annotations.

2. The positively evaluated co-occurrence results originate in many cases from texts that describe observations made in different contexts, i.e. different physiological or experimental conditions. Cell-cell relations assembled from co-occurrence results that combine different biological situations are not likely to exist in the organism.

A further observation is that in some cases different abstracts contained contrary statements regarding the same fact, which might be also caused by the different contexts in which the same signaling mechanisms have been investigated. A more detailed discussion of these problems as well as suggestions for improvements are presented in the next section.

6.3 Discussion

From the two kinds of data sources generally available, molecular databases and biomedical literature, the databases turned out to be too unspecific and incomplete in respect to the location sites of the molecules. Therefore, text mining is applied, and the previous results are restricted to the four cell types of interest in the context of neurodegenerative diseases.

The benefit of the restriction to a small set of interesting cell types is that all co-occurrence results that underly the hypotheses generation could be evaluated completely. Additionally, the initially high number of messenger-receptor interactions could be reduced by evaluating at first the relatively low numbers of co-occurrence results related to cell types. Thus, continuing the approach presented in the previous section, the application of further filter steps reduces the amount of co-occurrence hits to be evaluated and finally only hypotheses from positively evaluated co-occurrence hits are created.

Interestingly, it can be observed that the precision values of the co-occurrence results roughly show a pattern similar to that as it was found in the different searches for the complete list of cell types (Section 5.3): the search for ligand-receptor bindings (`msngr-rec-bword`) performs worst, the search for expression of receptors in cell types (`rec-cell-cword`) best, and the performance of the messenger release search (`cell-msngr-rword`) is in between. This underlines that the success of co-occurrence approaches depends on the searched relations and is difficult to rate in general.

Although about one-third of the evaluated hypotheses are rated as valid, a larger number seems to be implausible. The first reason for this is of technical nature: the concepts imported from the MeSH terms as well as from the annotations of the MEDLINE texts contain too many terms from different hierarchical levels, which was not checked in advance. The effect is that for instance from a sentence containing the terms “cytokine” and

"astrocyte" the conclusion "astrocytes secrete interleukin-6 (IL-6)" might be drawn since "cytokine" and "IL-6" are synonym regarding the term lists of MeSH and MEDLINE. To improve the quality of the concepts in the aligned ONDEX ontology probably these two sources should be left out in future or at least edited manually in order to assure that only substance names are included and class names avoided. This evaluation result sheds also light on the importance of a high-quality biological entity thesaurus. Dictionaries like the MeSH terms or the UMLS contain a large number of entities, but should be evaluated carefully in respect to the specific application context.

The second reason for the resulting false-positives is specific for the application case of neurodegenerative diseases: the respective cell types can occur in different tissues (e.g. in the brain or in the spinal cord) and are investigated under various conditions, as e.g the affected tissues might be healthy or not, in a developmental or an adult state or the experiments might be conducted *in vivo* or *in vitro*. For example, most of the autocrine signals inferred with this approach turned out to be probably false since they originate from *in vitro* experiments where a cell culture has been treated with a messenger substance, which subsequently induced the production of its respective receptor. From this fact one can not assume that these cells communicate with themselves *in vivo* by such a ligand-receptor combination, because the receptor might not be expressed in the cells under normal conditions.

Consequently, conclusions drawn from a single sentence of an abstract might hold true only under specific conditions, which can not be recognized any more by considering only this sentence. Thus, the restriction to sentences helped to find and to evaluate a reasonable amount of true-positive co-occurrences, but assembling these conclusions into cell-cell signaling hypotheses is ambiguous in the context especially of neurodegenerative diseases.

Whereas the first problem regarding the concept names could be resolved straightforwardly, the second problem is more complex. In order to include information on e.g., experimental or physiological conditions, the design of a subsequent approach should consider the context of a co-occurrence hit. Such information might be contained in the abstract or the full paper of the detected sentence. For this purpose, co-occurrence strategies might not be sufficient since relations between sentences have to be considered and therefore, full sentence parsers could be more appropriate to apply. However, since full sentence parsers are much more costly regarding computation time, an application of such tools would in many cases only be feasible by using an approach like the presented one as filter for appropriate texts.

Concluding this section, concept co-occurrence results for cell types relevant in neurodegenerative diseases are completely evaluated and cell-cell relation hypotheses generated. By applying several filter steps, the sets of resulting co-occurrences and hypotheses are of manageable size. Although the examination of the hypotheses turned out to be more complicated than expected, the retrieval of about one-third known cell-cell relations in the validation set is a reasonable result and comparable to other text mining work (Winnenburg, 2005) as well as to the 14% known relations in all relations extracted from the CSNDB (Section 4.2.4).

Chapter 7

Conclusions

The goal of this work was to reconstruct and analyze intercellular signaling networks, i.e. signaling relations between cells based on ligand-receptor interactions. Therefore, two generally different kinds of data sources were accessed: molecular databases and the biomedical literature. For both sources it turned out that explicit information on complete cell-cell signals is not available, but rather has to be combined from information on the different components of a cell signal, which are the messenger release from a cell, the binding of a messenger to a respective receptor on a target cell and the existence of this receptor in the target cell.

For these reasons are the reconstructed cell signals of hypothetical character, independently from the chosen data source. Validating hypotheses is difficult since even if they reflect the biological reality correctly, they might have not yet been investigated experimentally. This is additionally complicated by the usually large numbers of resulting hypotheses arising from even few combined cell signaling components. Consequently, the challenges to be resolved were the analysis and validation as well as the visualization of dense networks consisting of uncertain connections.

For this purpose, models capturing cell signals on different levels as well as corresponding graph representations were developed and applied to the data sources. With the proposed bi- and tripartite graphs the number of edges in signaling networks can be reduced and thus these are particularly of use in network visualization and support manual inspection of the resulting signals.

Regarding the available data sources, an examination of molecular databases and preliminary studies with a specific database revealed that the main problem with databases is that the locations of the molecules are rather unspecific, if defined at all. The selected database CSNDB contains only very few cell types and the semantic of the location fields is not defined properly. However, sample networks were reconstructed from the CSNDB by applying two reconstruction approaches. In the network resulting from the second approach, about 14% of the reconstructed cell-cell interactions were validated as true-positive. Statistical analysis was difficult to perform since the network is dense and consists mainly of only one giant strong component. In a sample application on the organ subnetwork a correlation between the signaling intensity and the physical distance of organs could not be found.

Therefore, instead of using inappropriate third party databases, the biomedical lit-

erature was accessed by applying a concept based text mining approach on MEDLINE abstracts. The advantage is that in this case the searched locations are defined as input of the reconstruction approach. In order to avoid simple text searches, the ONDEX framework has been developed in a collaborative work. With ONDEX, a pre-filtered set of texts is annotated by an ontology consisting of a number of imported and aligned databases and ontologies. This allows concept based approaches in text indexing, co-occurrence searches and hypotheses generation.

Applying concept based co-occurrence searches, a network of cell signals could be extracted and subsequently reduced by refining the co-occurrence searches. Manual evaluation of randomly selected samples showed that the precision depends on the searched signaling component and how this component is described in the texts. Due to stable expression in the texts, the existence of receptors in the target cells performed best with a precision of about 70%, whereas ligand-receptor bindings were most difficult to detect.

However, even with the final refinement of the resulting hypotheses there are still too many for an exhaustive evaluation. Therefore, both reconstruction results (i.e. from the sample database as well as from text) were inspected whether they can be applied on a set of four cell types important in the context of neurodegenerative diseases. The currently available databases could not be used here since the respective cell types are not present, whereas the respective concept based co-occurrence results from the text mining approach could be completely evaluated and a set of hypotheses based on these valid co-occurrences was generated.

For these hypotheses about one-third of a randomly selected sample could be shown to reflect known knowledge correctly, but the remaining hypotheses were rated as false-positives for two main reasons: firstly, some imported data sources caused synonym names in several concepts which should not be regarded as synonyms and secondly, the combination of positively evaluated co-occurrences often returns implausible cell-cell signals if they originate from different experimental settings or physiological conditions. Of the two problems, the second one was not easily solvable within the approach chosen here. To detect properly the context of a sentence it is probably most appropriate to apply tools that are able to reconstruct the grammar of a text and thus the relations of its sentences.

Hence, a benefit of the presented approach is the selection of texts likely containing the searched contents from millions of other texts with manageable effort. Thus, this strategy returns a small set of probably relevant texts that would have been hard to find by manual search queries. Also it enables the application of more sophisticated and therefore computationally expensive text mining tools. Furthermore, other applications in extracellular signaling might be not necessarily that strongly related to an experimental or physiological context and concept based co-occurrence searches are thus still reasonable to apply. Here, the different precision rates resulting for different co-occurrence searches indicate that some facts are better to extract by co-occurrence approaches than others.

Concluding from this, the concept based methodology of this thesis can be used to extract cell-cell signaling hypotheses that either can be applied directly or serve as base for more intensive text analysis approaches. Thus, concept based text extraction methodologies support the understanding of systems as complex as intercellular signaling.

Appendix

A Network extraction and text mining tools

In this section we introduce selected tools that implement one or more of the approaches discussed for each step of the network extraction workflow (Figure 2.5) in Section 2.3. Figure 7.1 gives an overview of recently developed and available software.

Examples for integrated applications that combine all steps of the workflow into one system are PIES (Wong, 2001), SUISEKI (Blaschke *et al.*, 2002), PreBIND (Donaldson *et al.*, 2003), GeneWays (Rzhetsky *et al.*, 2004) or PASTA (Gaizauskas *et al.*, 2003, tool no. (1) in Figure 7.1). The commercial software package PathwayAssist (2) also addresses the whole workflow. It uses MedScan (Novichkova *et al.*, 2003; Daraselia *et al.*, 2004) as module for textmining, which is also available separately and based on NLP techniques. After retrieving MEDLINE abstracts according to a user-defined query, sentences that do not contain at least one concept of a dictionary are filtered out. The remaining sentences are further processed with a syntactic parser and a semantic interpreter. The resulting relationships can then be visualised and analysed within PathwayAssist. The reported precision is 91% with a recall of 21%.

Chilibot (Chen and Sharp, 2004, tool no. 3) is a web service to construct networks from genes, proteins, drugs and other biological concepts. It uses the E-Utilities (4) service (ESearch and EFetch) at NCBI for retrieval of documents by submitting a query consisting of the pairwise combinations of the user's input terms and their synonyms. Acronyms contained in the user input are automatically resolved to their long-term phrases. Retrieved abstracts containing less than 30% of the acronym's phrase terms are rejected. Sentences from the abstracts that contain two or more query terms and synonyms are further processed by the POS tagger TnT (Brants, 2000, tool no. 5) and the shallow parser CASS (6). Following that, the resulting sentences are classified into one of six categories according to the presence/absence of terms indicating special relationships. For visualization of the extracted relationships AiSee (7) is used in Chilibot. The extracted network can in addition be used for navigating the related literature. The precision of the system was determined to be between 74% and 79% depending on the category and the recall to be about 90%.

PubGene (Jenssen *et al.*, 2001, tool no. 8) is an integrated system widely used in different projects. It is a commercial tool, but developed in academic research. The basic version described in Jenssen *et al.* (2001) uses a dictionary of gene symbols and names collected from HUGO nomenclature database, LocusLink, GDB and GENATLAS

Nb	Name	Main-WWW	Methods				Availability	Platforms
			T	E	R	N		
1	PASTA	http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/	x	x	x	x	Public	Web
2	PathwayAssist	http://www.ariadnegenomics.com/products/pathway.html	x	x	x	x	Commercial	Win
3	Chilibot	http://www.chilibot.net/	x	x	x	x	Public	Web
4	E-Utilities	http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html	x				Public	Web, Java
5	TnT	http://www.coli.uni-saarland.de/~thorsten/tnt/		x			Public	Unix
6	CASS	http://www.vinartus.com/spa/		x			Open Source	Unix
7	AiSee	http://www.aisee.com/				x	Commercial	Win, Unix
8	PubGene	http://www.pubgene.org/	x	x	x	x	Commercial	Web, Win
9	GraphViz	http://www.graphviz.org/				x	Open Source	Lin, Win
10	BioNLP	http://www.geneticxchange.com/		x	x		Commercial	Java
11	GATE	http://www.gate.ac.uk/		x	x		Open Source	Java
12	ONDEX	http://sourceforge.net/projects/ondex/	x	x	x	x	Open Source	Lin
13	MedlineR	http://dbsr.duke.edu/pub/MedlineR/	x	x	x		Open Source	R
14	Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/				x	Public	Win
15	PubMatrix	http://pubmatrix.grc.nia.nih.gov/	x	x	x		Public	Web
16	iHop	http://www.pdg.cnb.uam.es/UniPub/iHOP/	x	x	x	x	Public	Web
17	MedKit	http://metnetdb.gdcb.iastate.edu/medkit/	x				Open Source	Java
18	Textomy	http://www.litminer.ca/		x	x		n.a.	n.a.
19	Snowball	http://snowball.tartarus.org/		x			Open Source	Lin, Java
20	Qtag	http://www.english.bham.ac.uk/staff/omason/software/qtag.html		x			Different Lic.	Java
21	NLProt	http://cubic.bioc.columbia.edu/services/nlprot/index.html		x			Public	Lin, Win
22	Ingenuity	http://www.ingenuity.com				x	Commercial	Web
23	Cytoscape	http://cytoscape.org				x	Open Source	Java
24	Osprey	http://biodata.mshri.on.ca/osprey/				x	Different Lic.	Lin, Win

Figure 7.1: Available tools for network extraction and text mining. Only tools with maintained web sites and at least available at request are listed.

to identify genes in Medline. Each gene thereby is represented by its primary gene symbol. With the resulting gene article index co-occurrences of pairs of genes in the abstracts are calculated (see also "Relations" in the previous section). The retrieved network can be enriched with DNA microarray data. The visualization is done with GraphViz (9).

The systems described so far integrate all parts of the overall workflow. Building blocks of these applications are tools that cover either one task, e.g. TnT, or many parts, e.g. BioNLP (Ng and Wong, 1999, tool no. 10). A public available framework that provides the basic architecture for the development of information extraction applications is GATE (Cunningham *et al.*, 2002, tool no. 11). In the field of biological relation mining it is used e.g. in PASTA (Gaizauskas *et al.*, 2003) and by Karopka *et al.* (2004). GATE includes a set of components, which can be replaced or extended easily as the framework is provided as a Java API. Beside usual modules like a Tokenizer, a Sentence Splitter or a Tagger, components for recognizing relations and finding identical entities (Orthomatcher, Coreferencer) are available.

The ONDEX suite (12) is intended for integration of databases, network extraction and graph analysis (Section 5). Here, a concept based entity recognition using mapped ontologies is applied in a first step (see also Section 2.3.3) and used for text mining with a co-occurrence search. It is not restricted to Pubmed abstracts as texts are imported into a relational database format (PostgreSQL).

The library MedlineR (Lin *et al.*, 2004, tool no. 13) uses the statistical environment and programming language R to define procedures for retrieving articles from NCBI, mapping

terms to MeSH and mainly to calculate co-occurrences of terms. The visualization of the associations is realized through the generation of an output file in the Pajek (14) format.

PubMatrix (Becker *et al.*, 2003, tool no. 15) is in contrast to MedlineR a web-based tool intended for interactive querying. To calculate a co-occurrence matrix the user has to define two lists of terms, a search list and a modifier list. The terms of the list, which can be a simple keyword lists or gene symbols are used to create PubMed queries. This is realized by pairwise combining the terms of the different lists. Finally, the resulting matrix contains the frequency of co-occurrences. Another interactive querying tool is the iHOP service (16). It enables the search of genes in a pre-calculated co-occurrence network of genes and proteins (from eight organisms). In contrast to other systems the user retrieves fragments of sentences, which contain relations of the searched gene, and then selects relevant relations that should be added to a user specific literature network.

Finally, there exist a number of software packages that can be used in each single step of the network extraction workflow (Figure 2.5): The acquisition of texts can simply be done by using the E-Utilities of NCBI. MedKit (Ding and Berleant, 2005, tool no. 17) is also very useful for this purpose and more powerful. On the other hand more sophisticated methods can be applied to get more appropriate text corpora. Textomy (Donaldson *et al.*, 2003, tool no. 18), for example, is part of the PreBIND (Donaldson *et al.*, 2003) system and uses Support Vector Machines for classifying texts.

For identifying entities in text in most systems standard NLP techniques can be applied. In the biomedical domain public available tools have already been used e.g. Snowball (19) for stemming or Qtag (20) for part-of-speech tagging. Specialised taggers for biological knowledge also exist but might not be publicly available.

A publicly available system which addresses this task is NLProt (Mika and Rost, 2004, tool no. 21). It uses different dictionaries, e.g. a protein names dictionary extracted from Uniprot and a common names dictionary derived from Merriam-Webster, in combination with support vector machines (SVMs). For training the SVMs in the first step each abstract is split into single tokens separated by spaces. Out of this tokens sample phrases are constructed which are composed of a central part and a preceding respectively following environment. This enables the system to train the system for different purposes, e.g. one SVM was trained on central words and one for the environment. The system achieves a precision of 75% and a recall of 76% even for novel protein names.

Analysis and visualisation of the generated networks can be supported using specialised biological pathway and network analysis tools, as e.g. Ingenuity (22), Cytoscape (23), Osprey (24) or ONDEX (12). These tools enable users to analyse experimental data such as gene expression results in context of the biological networks. Ingenuity makes use of a knowledge base, but it could not be determined from the available information in the web whether this database or parts of it has been built using text mining.

But also more generic applications as for instance Pajek (14) are very useful especially in analyzing topological properties of the biological networks. For importing networks as text files the accepted formats of theses tools range from simple tab delimited files to common standards, as e.g. GML or PSI.

B Entities in the CSNDB

Locations

Table 7.1: 100 different locations that have been found by the first (column **I**) or the second (column **II**) CSNDB reconstruction approach. In the CSNDB the locations are listed in the fields `Tissue`, `Synthesis` or `Target` in the `Signal_Molecule` objects of a `Cell_Signaling`. The locations serve as nodes in the respective signaling networks. Here only locations are listed for which connections could be inferred. The last column (**Location type**) contains the manually assigned location type (body part, anatomical structure, organ, organ system, tissue, cell, cell part or embryo). The location type was not provided by the CSNDB. Location names in *italic* have been explicitly excluded in the second reconstruction approach. The order is alphabetically.

Location name	**I**	**II**	**Location type**
adipose tissue	x	x	tissue
adrenal cortex		x	tissue
adrenal gland	x	x	organ
adrenal medulla		x	tissue
aorta	x	x	organ
B lymphocyte	x	x	cell
basal ganglion	x	x	anatomical structure
basophil	x	x	cell
blood	x	x	organ
blood peripheral lymphocytes	x	x	cell
blood vessel	x	x	anatomical structure
bone	x	x	tissue
bone marrow	x	x	organ
brain	x	x	organ
breast	x	x	body part
cardiovascular system	x	x	organ system
cell line	x		removed in II
central nervous system	x	x	organ system
cerebellum	x	x	organ
cervix	x	x	organ
colon	x	x	organ
connective tissue	x	x	tissue
cortical axon	x	x	cell part
cytotoxic T cell	x	x	cell
dendrite	x	x	cell part
ear	x	x	body part
endothelium		x	tissue
epidermis		x	tissue

continued on next page

Location name	I	II	Location type
epithelial tissue		x	tissue
erythrocyte		x	cell
esophagus	x	x	organ
eye	x	x	body part
fetal brain	x		embryo, removed in II
fetal heart	x		embryo, removed in II
fiber	x	x	tissue
foreskin	x	x	anatomical structure
gall bladder	x	x	organ
gastrointestinal tract	x	x	organ system
genitourinary tract	x	x	organ system
germ cell	x	x	cell
glomerulus		x	tissue
GM progenitor cell		x	cell
head and neck	x	x	body part
heart	x	x	organ
helper T cell	x	x	cell
hippocampus	x	x	anatomical structure
hypophysis	x	x	organ
hypothalamus	x	x	anatomical structure
inhibitory synapse	x	x	cell part
islet of langerhans	x	x	anatomical structure
kidney	x	x	organ
larynx	x	x	organ
leucocyte		x	cell
liver	x	x	organ
lung	x	x	organ
lymph	x	x	organ
lymphocyte	x	x	cell
macrophage		x	cell
megakaryocyte	x	x	cell
monocyte	x	x	cell
mouth	x		body part
muscle	x	x	tissue
myocardium	x	x	anatomical structure
natural killer cell		x	cell
nervous system	x	x	organ system
neuron	x	x	cell

continued on next page

Location name	**I**	**II**	**Location type**
nose	x	x	body part
NT2 neuronal precursor	x	x	embryo
ovary	x	x	organ
pancreas	x	x	organ
parathyroid gland	x	x	organ
pDC2		x	cell
peripheral autonomic nervous system	x	x	organ system
peripheral nervous system	x	x	organ system
pineal body	x	x	organ
placenta	x	x	embryo
pooled	x		removed in II
postsynaptic neuron	x	x	cell
prostate	x	x	organ
retina	x	x	anatomical structure
skin	x	x	organ
smooth muscle	x	x	tissue
spinal cord	x	x	anatomical structure
spleen	x	x	organ
stomach	x	x	organ
suprachiasmatic nucleus	x	x	anatomical structure
synapse		x	cell part
synovial membrane	x	x	anatomical structure
T lymphocyte	x	x	cell
testis	x	x	organ
TH2		x	cell
thrombocyte	x	x	cell
thymus	x	x	organ
thyroid gland	x	x	organ
tongue	x	x	body part
tonsil	x	x	organ
ubiquitous	x		removed in II
uterus	x	x	organ
vascular smooth muscle		x	cell
whole embryo	x		embryo, removed in II

Selected `Cell_Signaling` objects in reconstruction approach I

Table 7.2: 74 `Cell_Signaling` objects selected in the first CSNDB reconstruction approach. In this approach the signalings are restricted to binary signalings of the type `ligand-receptor binding`, i.e. they consist of two molecules, one on the left ($\mathbf{M_1}$) and one on the right side ($\mathbf{M_2}$) of the signaling. Location links are inferred for the locations of $\mathbf{M_1}$ and $\mathbf{M_2}$. This table shows only the cell signalings that could be used for inferring location links, i.e. for both molecules are locations defined. The respective molecule types are liste in the second and the third column: hormone (H), cytokine (C), neurotransmitter (NT), receptor (R), ion channel (IC), transcription factor (TF) and enzyme (E). Multiple type assignments are possible as well as that no type is assigned (–). The order is alphabetically.

`Cell_Signaling`			$\mathbf{M_1}$	$\mathbf{M_1}$
adenosine	→	A2b receptor	NT	Rec
adrenomedullin	→	CRLR	H	Rec
anandamide	→	cannabinoid receptor	NT	Rec
anandamide	→	capsaicin receptor	NT	IC, Rec
anandamide	→	CB1	NT	Rec
ANP	→	ANP receptor	H	Enz, Rec
arginine vasopressin	→	V1a receptor	H	Rec
BLC	→	BLR-1	C	Rec
bombesin	→	bombesin receptor	NT	Rec
calcitonin	→	CRLR	H	Rec
CD40L	→	CD40	C	Rec
CGRP1	→	CRLR	H	Rec
CGRP2	→	CRLR	H	Rec
CNTF	→	CNTF receptor	C	Rec
cocaine	→	dopamine transporter	NT	Rec
delta9-THC	→	CB1	NT	Rec
EGF	→	EGF receptor	H	Enz, Rec
endomorphin-1	→	mu-opiate receptor	NT	Rec
endomorphin-2	→	mu-opiate receptor	NT	Rec
eotaxin	→	CCR3	C	Rec
Epo	→	Epo receptor	C	Rec
estradiol	→	estrogen receptor	H	Rec, TF
ethanol	→	NMDA receptor	–	IC, Rec
Fas ligand	→	DcR3	Rec	Rec
FGF1	→	FGFR1	H	Rec
FGF1	→	FGFR4	H	Rec
FGF2	→	FGFR1	H	Rec
FGF2	→	FGFR4	H	Rec
GABA	→	GABA-A receptor	NT	IC, Rec
GABA	→	GABA-B receptor	NT	IC, Rec
GABA	→	GABA-C receptor	NT	IC, Rec
gastrin-releasing peptide	→	GRP-R	NT	Rec
GCSF	→	GCSF receptor	C	Rec
GDNF	→	GDNF receptor	NT	Rec
GDNF	→	GDNFR-alpha	NT	Rec
GH	→	GH receptor	H	Rec

continued on next page

Cell_Signaling			$\mathbf{M_1}$	$\mathbf{M_1}$
glycine	→	GABA-A receptor	NT	IC, Rec
glycine	→	glycine receptor	NT	IC, Rec
GM-CSF	→	GM-CSF receptor	C	Rec
Gn-RH	→	Gn-RHR	H	Rec
IL-1	→	IL-1 receptor	C	Rec
IL-6	→	ErbB2	C	Rec
IL-6	→	ErbB3	C	Rec
insulin	→	insulin receptor	H	Rec
L-glutamate	→	AMPA receptor	NT	IC, Rec
L-glutamate	→	NMDA receptor	NT	IC, Rec
MCH	→	SLC-1	H, NT	Rec
MIP-1-beta	→	CCR5	C	Rec
morphine	→	mu-opiate receptor	–	Rec
motilin	→	motilin receptor	H	Rec
neuromedin B	→	NMB-R	NT	Rec
neurturin	→	NTNR-alpha	C	Rec
NGF	→	TrkA	H	Rec
NRG-2	→	ErbB3	NT	Rec
oxytocin	→	OTR	H	Rec
PEA	→	cannabinoid receptor	NT	Rec
PGE2	→	EP3 receptor	C	Rec
progesterone	→	OTR	H	Rec
progesterone	→	progesterone receptor	H	Rec, TF
PrRP	→	hGR3	H	Rec
SDF-1	→	CXCR4	C	Rec
semaphorin III	→	SemaIII receptor	–	Rec
serotonin	→	serotonin receptor	H, NT	Rec
substance P	→	substance P receptor	NT	Rec
testosterone	→	androgen receptor	H	Rec
thrombopoietin	→	thrombopoietin receptor	C	Rec
thrombopoietin agonist	→	thrombopoietin receptor	–	Rec
thyroxine	→	thyroxine receptor	H	Rec, TF
TNF-alpha	→	TNF receptor2	C	Rec
TRAIL	→	DcR1	C	Rec
TRAIL	→	DR4	C	Rec
TRAIL	→	DR5	C	Rec
TRAIL	→	TRID	C	Rec
urotensin-2	→	GPR14	H	Rec

Selected Cell_Signaling objects in reconstruction approach II

Table 7.3: 106 Cell_Signaling objects selected in the second CSNDB reconstruction approach. After the name of the signaling in the first column, the interaction type is given in the second column (lrb: ligand-receptor binding, ppi: protein-protein interaction, –: no type defined). The last two columns show the types of the two molecules selected from the cell signaling (H: hormone, C: cytokine, NT: neurotransmitter, Rec: receptor, Enz: enzyme, IC: ion channel, TF: transcription factor). If the cell signaling contains more than two molecules, the selected molecules are underlined, M_1 and M_2 appear in this order in the signaling. The signalings are ordered alphabetically and only such signalings are listed which could be used for inferring location links, i.e. all molecules have at least one location defined.

Cell_Signaling			Int	M_1	M_2
	→	TGF-beta1 + FKBP12	–	C	Enz, Rec
AA-NAT + serotonin	→	melatonin	–	H, NT	NT
acetylcholine	→	muscarinic acetylcholine receptor	–	NT	Rec
ACTH	→	ACTH receptor	–	H	Rec
adenosine	→	A2b receptor	lrb	NT	Rec
adrenaline	→	alpha2-adrenergic receptor	–	H, NT	Rec
adrenaline	→	beta-adrenergic receptor	–	H, NT	Rec
adrenomedullin	→	CRLR	lrb	H	Rec
anandamide	→	cannabinoid receptor	lrb	NT	Rec
anandamide	→	capsaicin receptor	lrb	NT	IC, Rec
anandamide	→	CB1	lrb	NT	Rec
angiotensin II	→	aldosterone	–	H	H
ANP	→	ANP receptor	lrb	H	Enz, Rec
arginine vasopressin	→	V1a receptor	lrb	H	Rec
BLC	→	BLR-1	lrb	C	Rec
bombesin	→	bombesin receptor	lrb	NT	Rec
calcitonin	→	CRLR	lrb	H	Rec
CD40L	→	CD40	lrb	C	Rec
CGRP1	→	CRLR	lrb	H	Rec
CGRP2	→	CRLR	lrb	H	Rec
CNTF	→	CNTF receptor	lrb	C	Rec
cocaine	→	dopamine transporter	lrb	NT	Rec
cortisol	→	glucocorticoid receptor	–	H	Rec, TF
CRH	→	ACTH	–	H, NT	H
delta9-THC	→	CB1	lrb	NT	Rec
desacetyl-alpha-melanocyte-stimulating hormone	→	MC4-R	–	H	Rec
digoxin	→	tetrodotoxin-sensitive Na(I) channel	–	H	IC
EGF	→	EGF receptor	lrb	H	Enz, Rec
endomorphin-1	→	mu-opiate receptor	lrb	NT	Rec
endomorphin-2	→	mu-opiate receptor	lrb	NT	Rec
eotaxin	→	CCR3	lrb	C	Rec
Epo	→	Epo receptor	lrb	C	Rec
estradiol	→	estrogen receptor	lrb	H	Rec, TF
estradiol	→	Maxi-K channel	lrb	H	IC
FGF1	→	FGFR1	lrb	H	Rec
FGF1	→	FGFR4	lrb	H	Rec
FGF2	→	FGFR1	lrb	H	Rec

continued on next page

Cell_Signaling			Int	M_1	M_2
FGF2	→	FGFR4	lrb	H	Rec
GABA	→	GABA-A receptor	lrb	NT	IC, Rec
GABA	→	GABA-B receptor	lrb	NT	IC, Rec
GABA	→	GABA-C receptor	lrb	NT	IC, Rec
gastrin-releasing peptide	→	GRP-R	lrb	NT	Rec
GCSF	→	GCSF receptor	lrb	C	Rec
GDNF	→	GDNF receptor	lrb	NT	Rec
GDNF	→	GDNFR-alpha	lrb	NT	Rec
GH	→	GH receptor	lrb	H	Rec
GH	→	IGF-1	–	H	H
GH-RH	→	GH	lrb	H	H
glycine	→	GABA-A receptor	lrb	NT	IC, Rec
glycine	→	glycine receptor	lrb	NT	IC, Rec
Gn-RH	→	FSH	–	H	H
Gn-RH	→	Gn-RHR	lrb	H	Rec
Gn-RH	→	LH	–	H	H
hGR3	→	prolactin	–	Rec	H
IL-1	→	IL-1 receptor	lrb	C	Rec
IL-1	→	IL-6	–	C	C
IL-12	→	IL-12 receptor	–	C	Rec
IL-12 receptor	→	IFN-gamma	–	Rec	C
IL-6	→	ErbB2	lrb	C	Rec
IL-6	→	ErbB3	lrb	C	Rec
insulin	→	insulin receptor	lrb	H	Rec
leptin	→	OB-RL	–	H	Rec
L-glutamate	→	AMPA receptor	lrb	NT	IC, Rec
L-glutamate	→	GluR5	lrb	NT	IC, Rec
L-glutamate	→	mGluR1	–	NT	IC, Rec
L-glutamate	→	NMDA receptor	lrb	NT	IC, Rec
LH	→	LH receptor	–	H	Rec
MCH	→	SLC-1	lrb	H, NT	Rec
MIP-1-beta	→	CCR5	lrb	C	Rec
motilin	→	motilin receptor	lrb	H	Rec
neuromedin B	→	NMB-R	lrb	NT	Rec
neurotrophin-3	→	TrkC	–	NT	Enz, Rec
neurturin	→	NTNR-alpha	lrb	C	Rec
NGF + TrkA	→	CREB	–	H	Rec
NGF	→	TrkA	lrb	H	Rec
NRG-2	→	ErbB3	lrb	NT	Rec
NRG-2	→	NMDA receptor	ppi	NT	IC, Rec
OPGL	→	OPG	lrb	H	Rec
OPGL	→	RANK	–	H	Rec
ouabain	→	tetrodotoxin-sensitive Na(I) channel	–	H	IC
oxytocin	→	OTR	lrb	H	Rec
PEA	→	cannabinoid receptor	lrb	NT	Rec

continued on next page

Cell_Signaling			Int	M_1	M_2
PGE2	→	EP3 receptor	lrb	C	Rec
PPAR-alpha	→	IL-1	–	Rec, TF	C
PRL-IH	→	prolactin	–	H	H
progesterone	→	OTR	lrb	H	Rec
progesterone	→	progesterone receptor	lrb	H	Rec, TF
PrRP	→	hGR3	lrb	H	Rec
renin	→	angiotensin II	–	H	H
SDF-1	→	CXCR4	lrb	C	Rec
serotonin	→	serotonin receptor	lrb	H, NT	Rec
somatostatin	→	GH	–	H, NT	H
substance P	→	substance P receptor	lrb	NT	Rec
Eta-1	↔	CD44	ppi	C	Rec
testosterone	→	androgen receptor	lrb	H	Rec
thrombopoietin	→	thrombopoietin receptor	lrb	C	Rec
thyroxine	→	thyroxine receptor	lrb	H	Rec, TF
TNF-alpha	→	CD44	–	C	Rec
TNF-alpha	→	TNF receptor2	lrb	C	Rec
TRAIL	→	DcR1	lrb	C	Rec
TRAIL	→	DR4	lrb	C	Rec
TRAIL	→	DR5	lrb	C	Rec
TRAIL	→	TRID	lrb	C	Rec
TRH	→	TSH	–	H	H
urotensin-2	→	GPR14	lrb	H	Rec
vitamin D	→	viatmin D receptor	lrb	H	Rec, TF

Selected ExtraCell_Signaling objects in reconstruction II

Table 7.4: ExtraCell_Signaling objects selected in the second CSNDB reconstruction approach. The CSNDB contains 15 ExtraCell_Signaling objects in total from which 8 have been selected since they contain information for intercellular signalings that could not be found in the previously checked Cell_Signaling and Gene_Expression objects. On each side of the ExtraCell_Signalings below the locations are given before the ":" (source and target on the left and on the right side respectively). On the right side of the ":" the name of the mediating ligand is given.

ExtraCell_Signaling		
hypophysis:FSH	→	ovary:FSH
ovary:estradiol	→	bone:estradiol
ovary:estradiol	→	breast:estradiol
ovary:estradiol	→	adipose tissue:estradiol
ovary:progesterone	→	breast:progesterone
ovary:progesterone	→	adipose tissue:progesterone
TH2:IL-4	→	DC1:IL-4
TH2:IL-4	→	pDC2:IL-4

C ONDEX implementation

In the following, all components of the ONDEX implementation are briefly described:

- PostgreSQL 7.4.1 (`http://www.postgresql.org/`): Although PostgreSQL includes also object oriented features, it was used only as standard relational DBMS. The SQL implementation of PostgreSQL conforms to the ANSI-SQL 92/99 standards. Additionally, PostgreSQL contains the TSearch2 tool for full text indexing (see below).

- Makefiles and shell scripts: Makefiles are currently the central interface for a user to start ONDEX processes. All Makefiles read a central configuration file containing global variables defining paths and other parameters. In turn the Makefiles may start shell scripts or SQL and Java programs. The most important process governed by Makefiles and Shell Scripts is the database installation: creation of the database scheme, creation of the import files from external database and text sources, actual import into the database, creation of database and indexes (Section 5.1.2, step 1 and Section 5.1.3, step 3). Makefiles and scripts take also care about the correct order of parsing and importing data. Shell scripts are especially used for text file manipulation and in case a process has to be started several times on a sequence of import files.

- Java 1.4.2 (`http://java.sun.com/`): Java is the language chosen for performing most of the ONDEX tasks at the core. These tasks comprise the import of data and texts (parsing of the flat files, applying specific rules depending on the data source, catching known interdependencies and syntactical errors in the sources; refer also to Section 5.1.2, step 1 and Section 5.1.3, step 3), the concept based indexing (Section 5.1.3, step 4) and the text mining methods (Section 5.1.3, step 5).

  The tools Snowball and QTag (see below) are accessed as Java libraries in order to add information to the generated import files. In case of indexing and text mining the database is accessed via JDBC. Depending on the amount of text and whether the text data is organized in one or several tables, different procedures have to be applied. Tests to perform the concept based indexing by using a ramdisk were also performed using Java. Furthermore, additional tools, as e.g. a program to access the MEDLINE web tools for filtering the abstracts according to a list of keywords, have been implemented with Java. And finally, the OVTK (Section 5.1.4) is completely implemented in Java.

- SQL (see also PostgreSQL 7.4.1, above): The main tasks performed with SQL are the actual import of data and text sources with the `COPY` command, the creation of database and full text indexes (see also TSearch2, below) and the ontology alignment (Section 5.1.2, step 2). Also, the identification of co-occurrences in the texts by searching the `IDENTIFIED_CONCEPT` table and the generation of hypotheses in the text mining part (Section 5.1.3, step 5), make use of SQL scripts.

- TSearch2 (`http://www.sai.msu.su/~megera/postgres/gist/tsearch/V2/`) is an integral component of the PostgreSQL DBMS. It creates a full text index on text columns of database tables. Using TSearch2 functions in `SELECT` statements improves the search performance essentially compared to a use of the `LIKE` operator of the `SELECT` command. Therefore, TSearch2 also applies the Snowball word stemmer (see below), which is the reason for using the same tool for other word stemming tasks in ONDEX.

  TSearch2 has been added with a new rank function to score the results of the concept based indexing regarding homonym detection and word sense disambiguation (Section 5.1.3, step 4). For this purpose, the ANSI C with the GNU C compiler has been used since PostgreSQL and TSearch2 are implemented in this language.

- Snowball (`http://snowball.tartarus.org/`): Snowball is a word stemming tool and implemented as Java library. It is used by the TSearch2, the full text indexer of PostgreSQL. To match concept names of the imported ontologies and databases correctly to words of text, which are indexed with TSearch2, also all ONDEX concept names are stemmed with Snowball. The table `CONCEPT_NAME` therefore has also an additional column `name_stemmed` containing the word stemmer results.

- QTag (`http://www.english.bham.ac.uk/staff/omason/software/qtag.html`): QTag is a program that reads text and for each token in the text returns the part-of-speech (e.g. noun, verb, punctuation, etc) by applying statistical methods. It is implemented in Java and used in both import procedures in ONDEX: for concept names when importing databases and ontologies (Section 5.1.2, step 1) and for importing texts (Section 5.1.3, step 3). So, also for this application of natural language processing (NLP) both sources are treated equally (correspondingly to the word stemming described before). Both concerned tables, `CONCEPT_NAME` and `TEXT` contain seperate columns, `name_stemmed` and `free_text_tagged` respectively, to store the results of word stemming and part-of-speech tagging.

D Text mining input lists

Entity lists

There are three different entity lists defined: for cell types (`cell`), first messengers (`msngr`) and receptors (`rec`). They are all manually collected from the MeSH terms (version from 2005), only for the cell type list additionally the CellOntology (CELL) was used (see also Table 5.1). The lists shown below are the complete input lists, but not all synonym names are given. All presented exemplary concept names are selected from the concept identified by the original ID of the respective ontology (MESH or CL) shown in the second and fourth column. Since these are the input lists, other synonyms from concepts of different ontologies might be mapped to these concepts subsequently. Also mappings between MESH and CL might be created and some concepts from the lists collapsed into one. In the concept based indexing process different spellings are applied. The lists are ordered by the ontolgy identifier (ID) from left to right and from top to bottom.

Table 7.5: Cell types collected from MeSH and the CellOntology (`cell`).

Cell type name(s)	ID	**Cell type name(s)**	ID
nurse cell	CL:0000026	monocyte stem cell, monoblast	CL:0000040
neutrophil stem cell, myeloblast	CL:0000042	cfu-gemm, colony forming unit granulocyte erythrocyte macrophage and megakaryocyte, multipotential myeloid stem cell, pluripotent stem cell	CL:0000049
cfu-em, megakaryocyte erythroid progenitor cell	CL:0000050	lymphopoietic stem cell	CL:0000051
enamel secreting cell	CL:0000053	bone matrix secreting cell	CL:0000054
ependymocyte, ependymal cell	CL:0000065	blood vessel endothelial cell	CL:0000071
epithelial cell of lung	CL:0000082	epithelial cell of pancreas	CL:0000083
germ line stem cell	CL:0000085	male germ line stem cell	CL:0000089
female germ line stem cell	CL:0000090	osteoclast	CL:0000092
osteochondroclast	CL:0000093	interneuron	CL:0000099
motor neuron	CL:0000100	sensory neuron	CL:0000101
polymodal neuron	CL:0000102	bipolar neuron	CL:0000103
multipolar neuron	CL:0000104	pseudounipolar neuron	CL:0000105
unipolar neuron	CL:0000106	autonomic neuron	CL:0000107
cholinergic neuron	CL:0000108	adrenergic neuron	CL:0000109
peptidergic neuron	CL:0000110	peripheral neuron	CL:0000111
columnar neuron	CL:0000112	mononuclear phagocyte	CL:0000113
basket cell	CL:0000118	Golgi cell	CL:0000119
granule cell	CL:0000120	stellate cell	CL:0000122
macroglial cell	CL:0000126	corneal endothelial cell	CL:0000132
fibrocyte	CL:0000135	fat cell, adipocyte, lipocyte	CL:0000136
odontocyte	CL:0000140	cementocyte	CL:0000141
hyalocyte	CL:0000142	simple columnar epithelial cell	CL:0000146
pigment cell	CL:0000147	Clara cell	CL:0000158
phaeochromocyte, chromaffin cell	CL:0000166	insulin secreting cell	CL:0000168

continued on next page

Cell type name(s)	ID	Cell type name(s)	ID
pancreatic B cell, beta cell	CL:0000169	glucagon secreting cell	CL:0000170
pancreatic A cell, alpha cell	CL:0000171	somatostatin secreting cell	CL:0000172
pancreatic D cell	CL:0000173	testosterone secreting cell	CL:0000177
progesterone secreting cell	CL:0000179	estradiol secreting cell	CL:0000180
myoepithelial cell	CL:0000185	myofibroblast cell	CL:0000186
skeletal muscle cell, striated muscle cell	CL:0000188	red muscle cell, slow muscle cell	CL:0000189
fast muscle cell, white muscle cell	CL:0000190	nodal cardiac cell	CL:0000194
Purkinje fiber	CL:0000195	thermoreceptor cell	CL:0000205
synovial cell	CL:0000214	lymphoblast	CL:0000229
T lymphoblast	CL:0000230	B lymphoblast	CL:0000231
pole cell	CL:0000301	IgM B lymphocyte	CL:0000302
IgG B lymphocyte	CL:0000303	IgA B lymphocyte	CL:0000304
IgE B lymphocyte	CL:0000305	epithelial cell of trachea, tracheocyte	CL:0000307
keratinocyte	CL:0000312	tears secreting cell	CL:0000315
sebum secreting cell	CL:0000317	sweat secreting cell	CL:0000318
glycogen accumulating cell	CL:0000326	adrenal medulla cell, chromafin cell	CL:0000336
choroidal cell	CL:0000348	sphincter associated smooth muscle cell	CL:0000358
vascular associated smooth muscle cell	CL:0000359	xanthophore	CL:0000430
iridiophore	CL:0000431	vaginal lubricant secreting cell	CL:0000436
delta basophil, follicle stimulating hormone secreting cell, gonadotroph	CL:0000437	luteinizing hormone secreting cell, gonadotroph	CL:0000438
mammotrophic cell, mammotroph, prolactin secreting cell	CL:0000439	melanocyte stimulating hormone secreting cell	CL:0000440
calcitonin secreting cell	CL:0000443	obliquely striated muscle cell	CL:0000444
parathyroid hormone secreting cell	CL:0000446	carbohydrate secreting cell	CL:0000447
white fat cell	CL:0000448	brown fat cell	CL:0000449
thyroid hormone secreting cell	CL:0000452	epinephrin secreting cell	CL:0000454
mineralocorticoid secreting cell	CL:0000456	glucocorticoid secreting cell	CL:0000460
Cardioblast	CL:0000465	adrenocorticotrophic hormone secreting cell. corticotroph	CL:0000467
growth hormone secreting cell, somatrophic cell, somatotroph	CL:0000471	pericardial cell	CL:0000474
beta-basophil, thyrotroph, thyroid stimulating hormone secreting cell	CL:0000476	follicle cell	CL:0000477
oxytocin stimulating hormone secreting cell	CL:0000478	vasopressin stimulating hormone secreting cell	CL:0000479

continued on next page

Cell type name(s)	ID	Cell type name(s)	ID
secretin stimulating hormone secreting cell	CL:0000480	cholecystokin stimulating hormone secreting cell	CL:0000481
bombesin stimulating hormone secreting cell	CL:0000483	connective tissue type mast cell	CL:0000484
mucosal type mast cell	CL:0000485	follicular epithelial cell	CL:0000500
gastrin secreting cell	CL:0000509	foam cell	CL:0000517
mononuclear cytotrophoblast cell	CL:0000523	syncytiotrophoblast cell	CL:0000525
afferent neuron	CL:0000526	efferent neuron	CL:0000527
nitrergic neuron	CL:0000528	CAP motoneuron	CL:0000532
primary motor neuron	CL:0000533	secondary motor neuron	CL:0000536
basophilic erythroblast	CL:0000549	polychromatophilic erythroblast	CL:0000550
unimodal nocireceptor	CL:0000551	orthochromatic erythroblast	CL:0000552
megakaryoblast, megakaryocyte progenitor cell	CL:0000553	gastrin stimulating hormone secreting cell	CL:0000554
brush cell, monodendritic cell	CL:0000555	cfu-gm, colony forming unit granulocyte macrophage, myeloid progenitor cell	CL:0000557
reticulocyte	CL:0000558	promonocyte	CL:0000559
band form neutrophil	CL:0000560	amacrine neuron, amacrine cell	CL:0000561
promyelocyte	CL:0000564	fat body cell	CL:0000565
apud cell	CL:0000568	C cell, parafollicular cell	CL:0000570
leucophore	CL:0000571	melanophore	CL:0000572
corneal epithelial cell	CL:0000575	border follicle cell	CL:0000579
neutrophilic myelocyte	CL:0000580	metamyelocyte	CL:0000582
null cell	CL:0000585	cold sensing thermoreceptor cell	CL:0000587
small luteal cell	CL:0000590	warmth sensing thermoreceptor cell	CL:0000591
large luteal cell	CL:0000592	androgen secreting cell	CL:0000593
pressoreceptor cell	CL:0000602	eosinophil stem cell	CL:0000611
eosinophilic myelocyte	CL:0000612	basophilic stem cell	CL:0000613
basophilic myelocyte	CL:0000614	GABAergic neuron	CL:0000617
acinar cell	CL:0000622	ito cell, perisinusoidal cell	CL:0000632
Hensen cell	CL:0000633	Claudius cell	CL:0000634
Deiter's cell, phalangeal cell	CL:0000635	Muller cell	CL:0000636
chromophobic cell	CL:0000641	folliculostellate cell, FS cell	CL:0000642
tanycyte	CL:0000643	Bergmann glial cell	CL:0000644
pituicyte	CL:0000645	juxtaglomerular cell	CL:0000648
prickle cell	CL:0000649	mesangial cell	CL:0000650
mucous neck cell — neck cell	CL:0000651	pinealocyte	CL:0000652
podocyte	CL:0000653	centripetally migrating follicle cell	CL:0000671
Ameloblast	D000565	Astroglias, Astrocyte	D001253
Basophils	D001491	Blastomeres, Blastocyte	D001757
Thrombocyte, Blood Platelets	D001792	Interdigitating Cell, Dendritic Cell, Veiled Cell	D003713

continued on next page

Cell type name(s)	ID	**Cell type name(s)**	ID
Enterochromaffin Cell, Argentaffin Cell	D004759	Eosinophils	D004804
Pronormoblast, Erythroblast, Normoblast	D004900	Red Blood Cell, Erythrocyte	D004912
Fibroblast	D005347	Granulocyte	D006098
Granulosa Cell	D006107	Hair Cell	D006198
Inner Hair Cell	D006199	Helper Cell, Helper T-Lymphocyte, Helper-Inducer T-Cell	D006377
Hematopoietic Stem Cell, Hematopoietic Progenitor Cell,Hematopoietic Colony-Forming Units	D006412	Histiocyte	D006644
Killer Cell, K Cell	D007693	Kupffer Cell	D007728
Labyrinth Supporting Cell	D007760	Langerhans Cell	D007801
Leukocyte, White Blood Cell	D007962	Mononuclear Leukocyte	D007963
Testicular Interstitial Cell, Leydig Cell	D007985	Granulosa-Luteal Cell, Granulosa-Lutein Cell, Theca-Lutein Cell, Small Luteal Cell	D008184
Lymphocyte, Lymphoid Cell	D008214	Macrophages	D008264
Tissue Basophils, Mast Cell	D008407	Megakaryocyte	D008533
Melanocyte	D008544	Monocyte	D009000
Motor Neuron	D009046	Gamma Motor Neuron, Gamma-Efferent Motor Neuron	D009047
Glia, Glial Cell, Neuroglial Cell	D009457	LE Cell, Neutrophils, Polymorphonuclear Leukocyte	D009504
Odontoblast	D009804	Perineuronal Satellite Oligodendroglia Cell, Perineuronal Satellite Oligodendrocyte	D009836
Ovocyte	D009865	Oogonia	D009867
Osteoblast	D010006	Odontoclasts, Cementoclasts, Osteoclasts	D010010
Osteocyte	D010011	Unfertilized Eggs, Ova	D010063
Gastric Parietal Cell, Oxyntic Cell	D010295	Phagocyte	D010586
Plasma Cell, Plasmacyte	D010950	Purkinje Cell	D011689
Reticulocyte	D012156	Retinal Ganglion Cell	D012165
Schwann Cell	D012583	Sertoli Cell	D012708
Spermatoblast, Spermatids	D013087	Spermiocyte, Spermatocyte	D013090
Spermatophores, Spermatogonias	D013093	Spermatozoa	D013094
Mother Cell, Progenitor Cell, Colony Forming Units	D013234	Suppressor-Effector-T-Lymphocyte, , Suppressor-Effector T-Cell, Suppressor Cell	D013490

continued on next page

Cell type name(s)	ID	Cell type name(s)	ID
T-Cell, T-Lymphocyte, Thymus-Dependent Lymphocyte	D013601	TC1 Cell,TC2 Cell,Cytotoxic T Lymphocyte, Cell-Mediated Lympholytic Cell	D013602
Ovarian Interstitial Cell, Theca Cell, Theca Interna, Theca Externa	D013799	Zygotes, Fertilized Egg, Fertilized Ovum	D015053
CD4-Positive T-Lymphocyte, T4 Lymphocyte, T4 Cell	D015496	Keratinocyte	D015603
Epithelioid Cell	D015622	Erythropoietic Stem Cell, Erythropoietic Progenitor Cell, BFU-Es, CFU-E, Erythroid Burst-Forming Units, Erythroid Colony-Forming Units	D015672
Foreign Body Giant Cell	D015743	Langhans-Type Giant Cell	D015744
Lymphokine-Activated Killer Cell, LAK Cell	D015979	B-Lymphocyte Subsets, B-Cell Subsets	D016175
Tumor-Derived Activated Cell, Tumor-Infiltrating Lymphocyte	D016246	Activated Killer Monocyte	D016260
Pulmonary Macrophages, Alveolar Macrophages	D016676	Suppressor-Inducer T-Cell, Suppressor-Inducer T-Lymphocyte	D017112
Microglia	D017628	Lipocyte, Fat Cell, Adipocyte	D017667
Peritoneal Macrophages	D017737	Rod Photoreceptors, Rods (Retina)	D017948
Cone Photoreceptors, Cones (Retina)	D017949	Pyramidal Cell	D017966
Olfactory Receptor Neuron	D018034	Vestibular Hair Cell	D018069
Outer Hair Cell	D018072	CD8 Positive Lymphocyte, CD8 Positive T-Lymphocyte, T8 Lymphocyte, T8 Cell	D018414
Th1 Cell	D018417	Th2 Cell	D018418
Merkel Cell, Merkel Receptors	D018862	Enterochromaffin-like Cell, ECL Cell	D019861
G Cell, Gastrin Cell	D019863	Somatostatin Cell, D Cell, Pancreatic delta Cell	D019864
Gastric Chief Cell	D019872	Paneth Cell	D019879
Chondroblast, Chondrocyte	D019902	Rouget Cell, Pericyte	D020286
Goblet Cell	D020397	Follicular Dendritic Cell	D020566
Enterocyte	D020895	Hepatocyte, Hepatic Cell, Liver Cell	D022781
Myeloid Stem Cell, Myeloid Progenitor Cell	D023461	Oncocyte, Askenazy Cell, Hurthle Cell, Oxyphil Cell	D024862
Amacrine Cell	D025042	Perineuronal Satellite Cell	D027161
Heart Muscle Cell, Cardiac Myocyte, Cardiac Muscle Cell	D032383	Cardiac Myoblast, Cardioblast	D032386

continued on next page

Cell type name(s)	ID	**Cell type name(s)**	ID
Smooth Muscle Cell, Smooth Muscle Myocyte	D032389	Smooth Muscle Myoblast	D032390
Precursor Muscle Cell, Embryonic Muscle Cell, Myoblast	D032446	Skeletal Myoblast	D032448
Skeletal Muscle Satellite Cell, Myogenic Satellite Cell	D032496	Totipotent Stem Cell	D039901
Multipotent Stem Cell	D039902	Pluripotent Stem Cell	D039904

Table 7.6: Messenger substance names collected from the MeSH terms (`msngr`).

Messenger name(s)	ID	**Messenger name(s)**	ID
Acetylcholine	D000109	Adenocard, Adenosine	D000241
Adenosine Triphosphate, ATP, Striadyne	D000255	Adrenocorticotrophic Hormone, Adrenocorticotropin, ACTH	D000324
Aldosterone	D000450	alpha-Melanocyte-Stimulating Hormone	D000521
Androstenedione	D000735	Epiandrosterone	D000738
Angiotensin I	D000803	Angiotensin II	D000804
Angiotensin III	D000805	Pitressin	D001127
L-Aspartic Acid, L-Aspartate	D001224	C-Fragment Endorphin	D001615
beta-Thromboglobulin	D001620	Bombesin	D001839
Bradykinin	D001920	Connecting Peptide, C Peptide	D002096
Calcitrin, Calcitonin	D002116	Carnosine	D002336
Cholecystokinin, Pancreozymin, Uropancreozymin	D002766	Colony Stimulating Factors, Myeloid Cell Growth Inducer, Macrophage Granulocyte Inducer	D003115
Corticosterone	D003345	ACTH-Releasing Factor, Corticotropin Releasing Hormone	D003346
Cortisone, Adreson	D003348	Cortexolone, Cortodoxone, Reichstein's Substance S	D003350
Cytokinins	D003583	Androstenolone, Dehydroepiandrosterone	D003687
Delta Sleep-Inducing Peptide, DSIP	D003701	Choloxin, D-Thyroxine	D003918
Dihydroprogesterone	D004092	Iodogorgoic Acid	D004105
Dopamine, Intropin	D004298	Leucine Enkephalin	D004743
Methionine Enkephalin	D004744	Enteroglucagon, Gut Glucagon, Oxyntomodulin	D004763
Epidermal Growth Factor, EGF	D004815	Adrenaline, Vaponefrin, Epinephrine	D004837
Epitestosterone, 17-alpha-Testosterone	D004845	Erythropoietin	D004921
Estetrol	D004953	Estradiol	D004958
Epiestriol, Ovestin	D004964	Estrone, Estrovarin, Folliculin	D004970
Etiocholanolone	D005043	Follicle-Stimulating Hormone, FSH	D005640
gamma Aminobutyric Acid, GABA, Aminalone	D005680	Glucose Dependent Insulin Releasing Peptide, Gastric-Inhibitory Polypeptide	D005749
Glucagon	D005934	Glycine Phosphate	D005998
Human Chorionic Gonadotropin, HCG	D006063	Histamine	D006632
Cortisol	D006854	Hydroxyestrones	D006894
17-alpha-Hydroxypregnenolone	D006907	Insulin, Velosulin, Humulin	D007328
Insulin-Like Growth Factor I	D007334	Insulin Like Growth Factor II	D007335
Interferon Type I	D007370	Interferon	D007372
Interleukin-1, IL-1, Lymphocyte Activating Factor	D007375	Interleukin-2, IL-2, Lymphocyte Mitogenic Factor	D007376

continued on next page

Messenger name(s)	ID	Messenger name(s)	ID
Interleukin-3, IL-3, Eosinophil Mast Cell Growth Factor	D007377	Interleukin	D007378
Iodotyrosine	D007470	Isopentenyladenosine	D007541
Leukocyte Migration Inhibitory Factor	D007961	Luteinizing Hormone, LH, Interstitial Cell Stimulating Hormone, ICSH	D007986
Luteinizing Hormone-Releasing Hormone, LHRH	D007987	Lipotropin	D008083
Lymphokines, Lymphocyte Mediators	D008222	Tumor Necrosis Factor beta, TNF-beta	D008233
Macrophage Migration Inhibition Factor	D008263	Melatonin	D008550
Motilin	D009037	Intermedins, Melanocyte-Stimulating Hormones, MSH	D009074
Melanostatin	D009075	Melanocyte-Stimulating Hormone Releasing Hormone, MSH Releasing Hormone	D009076
Atrial Natriuretic Peptide, ANP, Atrial Natriuretic Factor, ANF	D009320	Neuropeptide Y, Neuropeptide Tyrosine	D009478
Neurophysins	D009481	Neurotensin	D009496
Nitrogen Protoxide	D009569	Noradrenaline, Levarterenol, Norepinephrine	D009638
Octopamine, Norsynephrine, Norsympatol	D009655	Ocytocin, Syntocinon, Oxytocin, Pitocin	D010121
Pancreatic Polypeptide	D010191	Pro-Vasoactive Intestinal Peptide, Peptide Histidine Isoleucine	D010451
Gonadostatin, Pituitary Hormone Release Inhibiting Hormone	D010905	Pituitary Hormone-Releasing Hormones, Hypothalamic Releasing Factor	D010906
Choriomammotrophin	D010928	Platelet Factor 4, PF 4	D010978
Pregnancy-Associated beta-Plasma Protein,Trophoblast-Specific beta-1 Glycoprotein	D011268	Pregnanediol	D011276
Pregnenolone	D011284	Proopiomelanocortin, Proopiocortin	D011333
Progesterone, Pregnenedione	D011374	Proinsulin	D011384
Prolaction, PRL, Mammotropin, Pituitary Mammotropic Hormone, Pituitary Lactogenic Hormone	D011388	Prolactin Release-Inhibiting Hormone	D011389
Prolactin Releasing Hormone	D011390	Relaxin	D012065
Secretin	D012633	Serotonin	D012701
Cholecystokinin Octapeptide, CCK-OP	D012844	Insulin Like Growth Factors	D013002
Somatotropin Release-Inhibiting Hormone	D013004	Pituitary Growth Hormone	D013006

continued on next page

Messenger name(s)	ID	Messenger name(s)	ID
Growth Hormone Releasing Factor, GHRH	D013007	Anaprotin, Andractim, Androstanolone	D013196
Substance P	D013373	T-Cell Suppressive Factors	D013491
Sustanon	D013739	Cholecystokinin 4, CKK-4, Gastrin Tetrapeptide, Tetragastrin	D013758
Tetrahydrocortisol	D013760	Tetrahydrocortisone	D013761
Thymuline	D013935	Thymins	D013946
Thymosin	D013947	Thymus Hormones	D013951
Thyroid Stimulating Hormone, TSH	D013972	Thyrotropin Releasing Hormone	D013973
Thyroxin	D013974	Transfer Factor	D014165
trans-Retinoic Acid	D014212	T3 Thyroid Hormone, Cytomel	D014284
Reverse T3 Thyroid hormone	D014285	Tumor Necrosis Factor alpha, TNF-alpha	D014409
Tyramine	D014439	Vasoactive Intestinal Peptide, VIP	D014660
Vasotocin	D014668	18-Hydrocorticosterone	D015069
18-Hydroxydesoxycorticosterone	D015070	Androstane 3,17 diol	D015113
Neuromedin K, Neurokinin B	D015287	Neuromedin L, Neurokinin A, Substance K	D015288
Human Chorionic Gonadotropin alpha Subunit, HCG-alpha	D015292	Recombinant Interferon alpha-2a	D015380
Recombinant Interferon alpha-2b	D015381	Calcitonin Gene-Related Peptide	D015740
Neuroleukin	D015782	Monokines	D015846
Interleukin-4, IL-4, B Cell Stimulatory Factor-1	D015847	Interleukin-5, IL-5, Eosinophil Differentiation Factor, T-Cell-Replacing Factor, B Cell Growth Factor II	D015848
Interleukin-6, IL-6, Hepatocyte Stimulating Factor	D015850	Interleukin-7, IL-7, Lymphopoietin-1	D015851
Macrophage Colony-Stimulating Factor, M-CSF	D016173	Granulocyte Macrophage Colony Stimulating Factor, CSF-GM	D016178
Granulocyte Colony-Stimulating Factor, G-CSF	D016179	Cytokines	D016207
Interleukin-8, IL-8, Neutrophil Activating Peptide	D016209	Platelet Transforming Growth Factor	D016212
Macrophage Activating Factor	D016215	Hematopoietic Stem Cell Stimulators, Hematopoietic Cell Growth Factor, Hematopoietin	D016298
Interleukin-10, IL-10, Cytokine Synthesis Inhibitory Factor	D016753	Interferon alpha, Leukocyte Interferon	D016898
Interferon beta, Fibroblast Interferon	D016899	Interleukin-9, IL-9,T Cell Growth Factor P40	D016906
Sermorelin Acetate, Sermorelin	D017337	Interleukin-11, IL-11, Adipogenesis Inhibitory Factor	D017370

continued on next page

Messenger name(s)	ID	Messenger name(s)	ID
Interleukin-12, IL-12, Natural Killer Cell Stimulatory Factor	D018664	Glutamic Acid	D018698
Interleukin-13, IL-13	D018793	alpha Endorphin	D018822
gamma Endorphin	D018823	Chemokines, Chemotactic Cytokines, Intercrines	D018925
Monocyte Chemotactic and Activating Factor, Monocyte Chemotactic Protein-1, MCP-1	D018932	Monocyte Chemotactic Protein, Monocyte Chemoattractant Protein	D018945
T-Cell RANTES Protein	D018946	Human Chorionic Gonadotropin beta Subunit, HCG-beta	D018997
Mast Cell Growth Factor	D019089	Dehydroisoandrosterone Sulfate, DHA Sulfate	D019314
17-Hydroxyprogesterone	D019326	Human Growth Hormone, hGH	D019382
Macrophage Inflammatory Proteins	D019402	Interleukin-14, IL-14, High Molecular Weight-B-Cell Growth Factor	D019404
Macrophage Inflammatory Protein 1, Stem Cell Inhibitor	D019407	Interleukin-15, IL-15	D019409
Interleukin-16, IL-16, Lymphocyte Chemoattractant Factor	D019410	CC Chemokines, beta Chemokines	D019742
CXC Chemokine, alpha-Chemokines	D019743	C Chemokines, gamma Chemokines	D019744
beta Melanocyte Stimulating Hormone,beta-MSH	D019824	gamma Melanocyte Stimulating Hormone,gamma-MSH	D019825
FMRF amide	D019835	Gastrin-Releasing Peptide	D019886
PYY Peptide, Peptide YY	D019894	C Type Natriuretic Peptide	D020098
Interleukin-17, IL-17	D020381	Interleukin-18, IL-18, Interferon-gamma Inducing Factor	D020382
CX3C Chemokines	D020523	Recombinant Interferon alpha-2c	D020659
Leptin, Obese Protein	D020738	D Ala2 NMe Phe4 Gly ol Enkephalin	D020875
Bis-Pen-Enkephalin	D020881	beta Inhibin	D028322
FSH-Releasing Protein, Activin	D028341	Luteinizing Hormone beta Chain, LH-beta	D037101
FSH beta	D037201	Thyroid Stimulating Hormone beta Subunit, TSH beta	D037322

Table 7.7: Receptor names collected from the MeSH terms (`rec`).

Receptor name(s)	ID	**Receptor name(s)**	ID
Cyclic AMP Receptor Proteins	D002373	alpha Adrenergic Receptor	D011942
beta Adrenergic Receptor	D011943	Androgen Receptor	D011944
Angiotensin Receptors	D011945	B-Cell Antigen Receptor	D011947
CCK Receptors	D011949	Acetylcholine Receptors, ACh Receptors, Cholinergic Receptors	D011950
Concanavalin A Receptor	D011952	Cyclic AMP Receptor	D011953
Epidermal Growth Factor Receptor, EGF Receptor	D011958	Estradiol Receptor	D011959
Follicle Stimulating Hormone Receptor, FSH Receptor	D011962	GABA A Receptor, Diazepam Receptor	D011963
Corticoid Type II Receptors	D011965	Gonadotropin Releasing Hormone Receptor	D011966
Histamine H1 Receptor	D011969	Histamine H2 Receptor	D011970
Insulin Receptor	D011972	Luteinizing Hormone Receptor, LH Receptor	D011974
Muscarinic Acetylcholine Receptor, Muscarinic Receptor	D011976	Nicotinic Acetylcholine Receptor	D011978
Progestin Receptor, Progesterone Receptor	D011980	Prolactin Receptor	D011981
Prostaglandin Receptor	D011982	Serotonin Receptor	D011985
Somatotropin Receptor, Growth Hormone Receptor	D011986	Thyroid Stimulating Hormone Receptor, THS Receptor	D011989
IL-2 Receptor	D015375	Integrin alphaXbeta2	D016167
Integrin alphaLbeta2	D016169	Integrin alpha-M beta-2	D016177
IL-3 Receptor	D016185	Macrophage Colony Stimulating Factor Receptor, CSF-1 Receptor	D016186
CD116 Antigens, Granulocyte-Macrophage Colony-Stimulating Factor Receptor, GM-CSF Receptor	D016187	Granulocyte Colony-Stimulating Factor Receptor, G-CSF Receptor	D016188
NMDA Receptor	D016194	T-Cell Receptor gamma-delta	D016692
T-Cell Receptor alpha-beta	D016693	T-Cell Antigen Receptor-CD3 Complex	D017260
Dopamine D1 Receptor	D017447	Dopamine D2 Receptor	D017448
mu Opioid Receptors	D017450	CD 32 Antigens, Fc gamma Receptor	D017452
CD 23 Antigens, Immunoglobulin E Receptor	D017455	Albumin Receptor	D017457
Aldosterone Receptor	D017458	Atriopeptin Receptors, Atrial Natriuretic Peptides Receptor, ANP Receptors, Atrial Natriuretic Factor Receptor, ANF Receptor	D017461

continued on next page

Receptor name(s)	ID	**Receptor name(s)**	ID
CD35 Antigens	D017463	CD 21 Antigens	D017464
delta Opioid Receptor	D017465	Endothelin Receptors	D017466
Erythropoietin Receptor	D017467	Fibroblast Growth Factor Receptor, FGF Receptor	D017468
Glutamate Receptor, Excitatory Amino Acid Receptor	D017470	Interferon Receptor	D017471
IL-1 Receptor	D017472	kappa Opioid Receptor	D017473
Neuropeptide Y Receptor	D017476	Platelet Derived Growth Factor Receptor, PDGF Receptor	D017479
sigma Opioid Receptor	D017480	Somatostatin Receptor	D017481
Thromboxanes Receptors, TP Receptor	D017482	Vasopressin Receptor	D017483
Insulin-Like Growth Factor Type 1 Receptor, IGF-I Receptor	D017526	Insulin-Like Growth Factor Type 2 Receptor, IGF-II Receptor	D017527
Bradykinin Receptor	D018002	Calcitonin Receptor	D018003
Neuromedin B Receptor, Gastrin Releasing Peptide Receptor	D018004	Vasoactive Intestinal Peptide Receptor, VIP Receptor	D018005
Glycine Receptor	D018009	Calcitonin Gene-Related Peptide Receptor, CGRP Receptor	D018015
Parathyroid Hormone Receptors	D018016	Corticotropin Releasing-Hormone Receptor, CRH Receptor	D018019
Thyrotropin Releasing Hormone Receptor, TRH Receptor	D018025	Glucagon Receptor	D018027
Neurotensin Receptor	D018028	Olfactory Receptor Proteins	D018035
Neurokinin-1 Receptor	D018040	Neurokinin-2 Receptor	D018041
Neurokinin-3 Receptor	D018042	Adrenocorticotropic Hormone Receptor, ACTH Receptor, Corticotropin Receptor	D018043
Oxytocin Receptor	D018045	Purinergic P1 Receptor, Adenosine Receptor	D018047
Purinergic P2 Receptor, ATP Receptor	D018048	Leukotriene Receptor, SRS-A Receptor	D018077
Prostaglandin E Receptor	D018078	GABA-B Receptor, Baclofen Receptor	D018080
AMPA Receptor, Quisqualate Receptor	D018091	Kainic Acid Receptor, Kainate Receptor	D018092
Metabotropic Glutamate Receptor	D018094	Histamine H3 Receptor	D018100
Leukotriene B4 Receptor	D018102	CD28 Antigens	D018106
Transforming Growth Factor beta Receptor, TGF-beta Receptor	D018125	Corticoid Type I Receptors	D018161
Vitamin D Receptor	D018167	Retinoic Acid Receptor	D018168
Thrombomodulin	D018180	TCDD Receptor, Polyaromatic Hydrocarbon Receptor	D018336

continued on next page

Receptor name(s)	ID	Receptor name(s)	ID
alpha-1 Adrenergic Receptor	D018340	alpha-2 Adrenergic Receptor	D018341
beta-1 Adrenergic Receptor	D018342	beta-2 Adrenergic Receptor	D018343
erbB-2 Receptors	D018719	CD18 Antigens, beta2 Integrin	D018821
CD14 Antigens, Lipopolysaccharide Receptor	D018950	CD36 Antigens, Thrombospondin Receptor	D018955
CD44 Antigens, Hyaluronan Receptor	D018960	CD117 Antigens, Stem Cell Factor Receptor, SCF Receptor	D019009
CD29 Antigens, beta1 Integrin	D019012	CD 95 Antigens, fas Receptor	D019014
CD42d Antigens, Platelet Glycoprotein GPIb IX Complex	D019038	GPIIb-IIIa Receptor	D019039
CD62L Antigens, Leukocyte Adhesion Molecule LAM-1	D019041	Polymeric Immunoglobulin Receptor	D019056
CC Chemokine Receptor 5	D019713	CXCR4 Receptor	D019718
Calcium Ryanodine Receptor Complex, Ryanodine Receptor	D019837	Hepatocyte Growth Factor Receptor, HGF Receptor	D019859
IL-6 Receptor	D019947	IL-4 Receptor	D019948
IL-7 Receptor	D020395	Platelet Derived Growth Factor alpha Receptor, PDGF alpha Receptor	D020796
Nerve Growth Factor Receptor	D020800	Ciliary Neurotrophic Factor Receptor	D020801
Neurotrophin 3 Receptor	D020812	trkB Receptor	D020813
erbB-3 Receptors	D020893	trkA Receptor	D020917
Cyclosporin Binding Protein	D021983	Tacrolimus Binding Protein 1A, Macrophilin-12	D022061
beta-3 Adrenergic Receptor	D022702	IL-8A Receptor	D023062
IL-8B Receptor	D023063	Activin Receptors Type I	D030201
Activin Receptors Type II	D030301	N-Acetylglucosamine Receptor	D034781
EphA1 Receptor	D036082	EphA2 Receptor	D036104
EphA3 Receptor	D036121	EphA4 Receptor	D036122
EphA5 Receptor	D036123	EphA6 Receptor	D036124
EphA7 Receptor	D036141	EphA8 Receptor	D036143
EphB2 Receptor	D036183	EphB5 Receptor	D036201
EphB3 Receptor	D036223	EphB4 Receptor	D036224
EphB1 Receptor	D036225	c erb A Protein	D037021
Thyroid Hormone Receptors beta, TR beta	D037042	Hepatic Asialoglycoprotein Receptor, Hepatic Lectin, Liver Carbohydrate Binding Protein	D037263
Integrin alpha2beta1	D038982	Integrin alpha4beta1	D039041
Integrin alpha5beta1	D039081	Integrin alpha6beta1	D039121
Integrin alpha6beta4	D039161	Integrin alpha3beta1	D039201
Integrin alpha1beta1	D039222	Integrin alphaVbeta3	D039302
Integrin alpha2	D039421	Integrin alpha3	D039422
Integrin alpha1	D039423	Integrin alpha4	D039441
Integrin alphaM	D039481	Integrin alpha5	D039482

continued on next page

Receptor name(s)	ID	**Receptor name(s)**	ID
Integrin alpha6	D039503	Integrin alphaX	D039521
Integrin alphaV	D039564	Integrin beta3	D039661
Integrin beta4	D039663	Semaphorin III Receptor	D039942
Neuropilin 2	D039943	Integrin alpha IIb	D040201
Endothelial Growth Factor Receptor	D040262	Vascular Endothelial Growth Factor Receptor-1	D040281
Vascular Endothelial Growth Factor Receptor-2	D040301	Vascular Endothelial Growth Factor Receptor-3	D040321
Integrin alphaL	D040881		

Word lists

All words in these lists are used for indexing the MEDLINE texts and selected from the WordNet dictionary (see Fellbaum (1998) and `wordnet.princeton.edu`). In each table, the second and the third columns show the part-of-speech (POS) and the WordNet identifier (WN ID) respectively. Equal words with different identifiers might be chosen if the meaning (in WordNet) is different.

Table 7.8: Words that indicate the *release* of substances or molecules from a cell (`rword` list).

Word	POS	WN ID	Word	POS	WN ID
produce	verb	01575778	prodcution	noun	00859333
release	noun	12785461	release, secrete	verb	00067106
secretion	noun	05095511	secretion	noun	12789685
segregate	verb	00481209	set free	verb	02422490
synthesize	verb	00623503	synthesis	noun	12800655
unleash	verb	01433584			

Table 7.9: Words that indicate the *binding* or interaction of substances or molecules (`bword` list).

Word	POS	WN ID	Word	POS	WN ID
adhere	verb	01316841	associate	verb	00689759
bind	verb	00551780	binding	noun	04494716
bond	noun	10697642	interact	verb	02305904
interaction	noun	10773922			

Table 7.10: Words that indicate that a cell *contains* or expresses specific receptor molecules (`cword` list).

Word	POS	WN ID	Word	POS	WN ID
contain	verb	02551275	contain	verb	02619957
express	verb	02082567	expression	noun	12715700
incorporate	verb	02551275	incorporation	noun	05421017
insert	verb	00181348	insertion	noun	00306609
internalization	noun	05421017	trafficking	verb	02195477

E Text mining results

Validated co-occurrence hits for neurodegenerative diseases

In the following the positively evaluated results of the co-occurrence searches are presented in abbreviated form. That means, only one substance name or its abbreviation (first messenger or receptor) is listed (although the hit might be resulting from a synonym name) and one example sentence where this co-occurrence has been found. The hit sentence can be found with the PMID number at the PubMed web interface to the MEDLINE database (`www.ncbi.nlm.nih.gov/entrez`). The sentence numbering starts with the title, i.e. sentence number three, for example, is the second sentence in the text of the abstract.

Table 7.11: True-positive `cell-msngr-rword` co-occurrence results in neurodegenerative diseases, i.e. all positive hits with first messenger substances that can be produced or secreted by the four relevant cell types (motor neurons, astrocytes, microglia and Schwann cells). The results are grouped by these cell types and sorted inside the four groups alphabetically by the first messenger name.

Messenger name	PMID	Sentence
Motor Neurons		
Acetylcholine	12482724	3
ACTH	9190120	8
Adenosine	12620503	2
ATP	8027518	2
Calcitonin Gene-Related Peptide	7966725	9
Chemokines	11549718	10
Dopamine	7301201	3
FMRFamide	12917365	2
Gastrin-Releasing Peptide	7907863	16
Glucagon	7301201	3
Glutamate	3249604	1
Insulin	1967914	3
POMC	7984050	8
Retinoic Acid	10357892	4
Substance P	2409171	2
Astrocytes		
ACTH	9651548	10
Adenosine	1681548	2
alpha IFN	2388040	2
Androstene	9927319	8
Angiotensin II	10646512	5
ANP	12629160	10
beta IFN	9282915	1
C-C Chemokines	9203678	1
Chemokines	12586731	2
Colony-Stimulating Factor	1382099	1
CXC Chemokines	11694326	6
Cytokines	9482215	3
DHEA	10593612	1
Erythropoietin	12380959	3

continued on next page

Messenger name	PMID	Sentence
Estradiol	10947808	2
GABA	9436790	4
Gastrin-Releasing Peptide	9006974	5
G-CSF	1382099	5
Glutamate	12235140	8
Glycine	2593181	7
GM-CSF	10421784	11
IGF	9037485	1
IGF-1	11549714	5
IL-1 beta	11286158	4
IL-2	1353061	4
IL-3	8087421	10
IL-6	8724985	1
IL-8	10580798	4
IL-12	9973393	4
IL-15	12572774	7
IL-18	10101231	10
Inflammatory Proteins MIP-2	11948246	1
Interferon	8926042	3
Interleukins	2463998	7
Lymphokines	8478988	2
M-CSF	2151455	1
Met-Enkephalin	8413825	9
MIP-1	9671961	8
Monocyte Chemotactic Protein-1	14722715	1
Monocyte Chemotactic Protein-1	14720224	3
Neuropeptide Y	11277982	2
Progestorone	14614261	1
Substance P	9802422	5
TGF-beta	12888549	8
Thymosin	11008214	9
TNF-alpha	11080817	5
Microglia		
alpha MSH	9620667	6
ATP	15129165	6
beta Endorphin	8112823	1
C-C Chemokines	10642753	4
Chemokines	12112366	3
CXC Chemokines	15312171	7
Cytokines	10374812	10
EGF	9283823	13
Glutamate	14715932	10
Histamine	11403935	1
IGF-I	15076729	2
IL-1	1874973	8

continued on next page

Messenger name	PMID	Sentence
IL-2	10375739	2
IL-3	9383038	9
IL-4	8377948	1
IL-5	7702706	8
IL-8	10950803	4
IL-10	11137576	1
IL-12	11356009	1
IL-15	8804052	7
IL-16	15175077	5
IL-18	10101231	12
Interleukins	15019947	1
Lymphokines	11377701	5
Macrophage Inflammatory Protein-1	11449364	5
Macrophage Inflammatory Proteins	10950803	4
M-CSF	8070891	10
Monocyte Chemotactic Protein-1	15139008	4
Monocyte Chemotactic Proteins	10950803	4
Monokines	12909303	6
Substance P	11113362	5
TGF-beta	2254955	4
Thymosin	7798904	8
TNF-alpha	7932874	11
Schwann Cells		
Acetylcholine	7463090	1
ATP	9483546	1
Chemoattractant Protein-1	11168559	9
Chemokines	11168559	3
c-kit Ligand	14679180	9
Cytokines	8982104	4
Dihydroprogesterone	11534984	3
Glutamate	9483546	6
IL-1	1894731	1
IL-6	10586289	1
IL-8	10580813	3
IL-12	8529135	9
Lymphokines	2146529	5
Macrophage Inflammatory Protein-1	15139590	8
Macrophage Migration Inhibitory Factor	12297465	1
Pregnenolone	14670648	12
Progesterone	8743966	8
TGF-beta	8457871	6
TNF-alpha	12953261	7

Table 7.12: True-positive `msngr-rec-bword` co-occurrence results in neurodegenerative diseases, i.e. all positive hits of interactions or bindings of first messenger substances and receptors. The results are sorted alphabetically by the first messenger name.

Messenger name	**Receptor name**	PMID	Sentence
Acetylcholine	Cholinergic Receptor	7855204	2
Acetylcholine	Muscarinic Receptors	10101037	2
Acetylcholine	Nicotinic Acetylcholine Receptor	7527881	3
ACTH	CRF Receptor	7477349	6
Adenosine	Adenosine Receptor	11082113	16
Adenosine	P2 Purinoceptors	8954905	1
alpha Chemokines	CXCR4 Receptor	10200343	1
Androstenedione	Androgen Receptor	3488063	2
ANF	ANF Receptor	1420611	3
Angiotensin II	Angiotensin II Receptor	7656287	3
ATP	P1 Purinoceptors	2847203	9
ATP	ATP Receptor	8840398	3
beta-Endorphin	mu Opioid Receptor	10854259	10
beta-Endorphin	delta Opioid Receptor	12670304	4
beta-Endorphin	kappa Opioid Receptor	2908136	2
Calcitonin Gene Related Peptide	Calcitonin Receptor	2828211	1
Chemokines	LFA-1	10613446	5
Chemokines	Integrin beta1	11989791	5
Chemokines	CXCR4 Receptor	11575704	4
Chemokines	IL-8 Receptor	7929358	2
Colony Stimulating Factor	CSF-1 Receptor	8384358	23
CSF-1	CSF-1 Receptor	2551961	9
CXC Chemokines	CXCR2 Receptor	10747307	3
Cytokines	IL-1 Receptor	10852706	5
Cytokines	PDGF Receptor	7546776	9
Cytokines	Adenosine Receptor	14530318	12
Cytokines	Retinoic Acid Receptor	9607817	7
Cytokines	alpha-2 Adrenergic Receptor	10808050	2
Cytokines	Hyaluronan Receptor	10652271	2
Cytokines	fas Receptor	12504821	2
Cytokines	HGF Receptor	12594808	2
Cytokines	IL-6 Receptor	9505191	5
Cytokines	IL-4 Receptor	10691892	5
Cytokines	PDGF alpha Receptor	7546776	9
Cytokines	CNTF Receptor	15051883	1
Cytokines	Integrin alphaV	11245625	3
Cytokines	Integrin beta3	9786457	5
Cytokines	VEGF Receptor	12872364	8
Cytokines	VEGF Receptor Type 2	12706123	5
DHEA	Androgen Receptor	9806358	5
Dopamine	Dopamine-D2 Receptor	11280926	3
EGF	EGF Receptor	7883816	12
GABA	GABA-A Receptor	6547630	7
GABA	GABA-B Receptor	9872315	2
Gastrin-Releasing Peptide	Bombesin Receptor	8788416	2

continued on next page

Messenger name	**Receptor name**	PMID	Sentence
Glutamate	NMDA Receptor	9504387	9
Glutamate	Glutamate Receptor	11173982	4
Glutamate	AMPA Receptor	8391661	2
Glutamate	Kainate Receptor	9630393	8
Glutamate	Metabotropic Glutamate Receptor	8189254	5
Glycine	NMDA Receptor	2163119	1
Glycine	Glycine Receptor	3023812	1
Glycine	AMPA Receptor	11855983	10
Histamine	Histamine H1 Receptor	1970573	3
IL-15	LFA-1	9502767	5
IL-18	IL-1 Receptor	9620656	6
IL-2	delta Opioid Receptor	9051743	5
IL-3	IL-3 Receptor	8943237	2
IL-6	Androgen Receptor	15129430	8
IL-6	IL-6 Receptor	9118960	7
IL-8	EGF Receptor	10702246	8
IL-8	CXCR2 Receptor	12548717	12
Insulin	Insulin Receptor	288716	2
Insulin Like Growth Factor	Insulin Receptor	499074	8
Insulin Like Growth Factor	IGF II Receptor	14523643	3
Insulin Like Growth Factor	EGF Receptor	6973821	3
Met-Enkephalin	mu Opioid Receptor	3032015	3
Met-Enkephalin	delta Opioid Receptor	1357608	5
Met-Enkephalin	kappa Opioid Receptor	6258931	1
Progesterone	Androgen Receptor	6706245	2
Progesterone	Glucocorticoid Receptor	1000505	8
Progesterone	Muscarinic Receptor	3374756	10
Progesterone	Mineralocorticoids Receptor	7575603	2
Retonic Acid	IGF-II Receptor	9811861	2
Retonic Acid	Retonic Acid Receptors	8387213	2
Substance P	Nicotinic Acetylcholine Receptor	7514262	3
Substance P	Bombesin Receptors	1383741	5
Substance P	NK-1 Receptor	12388097	3
TGF-beta	IGF-2 Receptor	10508563	2
TGF-beta	TGF-beta Receptor	2873833	9
TGF-beta	Integrin alphaV	10025398	1
Vasoactive Intestinal Polypeptide	delta Opioid Receptor	15126111	2

Table 7.13: True-positive `rec-cell-cword` co-occurrence results in neurodegenerative diseases, i.e. all positive hits with receptors that are contained in or expressed by the four relevant cell types (motor neurons, astrocytes, microglia and Schwann cells). The results are grouped by these cell types and sorted inside the four groups alphabetically by the first messenger name.

Receptor name	PMID	Sentence
Motor Neurons		
5 HT Receptor	12670306	12
ACh Receptor	7472435	11
AMPA Receptor	11279366	9
Androgen Receptor	10587588	1
BDNF Receptor	8083736	4
Benzodiazepine Receptor	9364456	1
beta4 Integrin	11064368	4
CGRP Receptor	2848610	11
CNTF Receptor	8945760	1
Endothelin Receptor	12941473	1
Eph Receptor	10673322	1
Eph-A4 Receptor	10646798	3
Excitatory Amino Acid Receptor	8895864	9
Glycine Receptor	2555150	7
HGF Receptor	10725250	9
IL-1 Receptor	11311987	2
IL-6 Receptor	9063729	5
Kainate Receptor	12429586	1
Metabotropic Glutamate Receptor	10982465	10
mu Opioid Receptor	7790855	8
Neurokinin-3 Receptor	11958875	1
Neuropilin-1	15094469	4
Neuropilin-2	15094469	5
NGF Receptor	8174770	3
Nicotinic Acetylcholine Receptor	2176713	2
NMDA Receptor	10886684	7
TGF beta Receptor	10842018	7
TRH Receptor	1280790	5
trkC Receptor	9822749	7
Vasopressin Receptor	10407169	10
Astrocytes		
ACh Receptor	7969896	6
Adenosine Receptor	9650577	2
Adrenergic alpha Receptor	2148555	10
Adrenergic beta Receptor	10559386	1
alpha-2 Adrenergic Receptor	6119369	2
alphav Integrin	11461157	2
AMPA Receptor	9405512	3
Androgen Receptor	11226751	3
Angiotensin II Receptor	1860709	8
ANP Receptor	1317098	3
Benzodiazepine Receptor	12106778	6
beta-2 Adrenergic Receptor	8723842	2

continued on next page

Receptor name	PMID	Sentence
beta3 Integrin	11470407	7
beta4 Integrin	8786405	5
Bombesin Receptor	9458349	3
Bradykinin Receptor	1338944	11
Calcitonin Receptor	12898703	9
CD44 Antigen	8355030	4
c-erbB-2 Protein	7826981	12
CRF Receptor	12898703	9
delta Receptor	9482211	13
Dopamine D2 Receptor	8057777	3
EGF Receptor	1468600	11
Endothelin Receptor	10799769	8
Eph Receptor	12944508	3
fas Receptor	12676530	9
Fc gamma Receptor	1386416	10
FGF Receptor	8793862	7
Flk-1	11934468	8
GABA-B Receptor	14550781	6
Glucocorticoid Receptor	10715588	5
Glutamate Receptor	2540340	11
H1 Receptor	1675832	2
IGF-II Receptor	1319501	10
IL-4 Receptor	11052816	2
IL-8Rbeta	10785334	8
Insulin Receptor	1851850	9
Integrin beta1	15042583	8
Kainate Receptor	9145303	1
LFA-1	7572280	4
Metabotropic Glutamate Receptor	10533045	2
Mineralocorticoid Receptor	7825881	2
Muscarinic Receptor	10413035	7
Neuropilin-1	15233640	12
Neurotensin Receptor	10625058	2
NGF Receptor	10446331	1
Nicotinic Acetylcholine Receptor	14681929	1
Opioid Receptor	15217373	6
Oxytocin Receptor	11754214	1
P2 Purinoceptor	8895885	1
PDGF alpha Receptor	8982160	3
PDGF Receptor	10559409	4
PGE Receptor	1324890	11
Retinoic Acid Receptor	11483254	1
Serotonin Receptor	8856328	1
Somatostatin Receptor	7595483	1
TGF beta Receptor	1320057	8

continued on next page

Receptor name	PMID	Sentence
Thrombomodulin	8969798	14
TR beta	7827337	3
TSH Receptor	7882998	2
Vasopressin Receptor	9630527	42
VEGF Receptor	9875268	6
VLA-1	1356158	9
VLA-2	1356158	9
VLA-6	1356158	9
Microglia		
alpha6beta1 Integrin	11880486	7
AMPA Recepto	15139014	5
Benzodiazepine Receptor	9483537	10
beta2 Integrin	14500997	5
beta-Adrenergic Receptor	12271472	2
Bradykinin Receptor	12551746	1
CR3 Receptor	1506289	10
CRH Receptor	12485415	2
CXCR4 Receptor	9218610	2
EGF Receptor	1883522	3
GABA-B Receptor	15019947	1
HGF Receptor	8380919	5
IL-3 Receptor	7643220	1
IL-4 Receptor	8071435	2
IL-6 Receptor	10861795	8
kappa Opioid Receptor	8755601	3
LFA-1	7533208	8
M-CSF Receptor	11520119	13
Metabotropic Glutamate Receptor	12358765	3
mu Opioid Receptor	9152411	4
Muscarinic Receptor	9839720	6
Neurokinin-1 Receptor	11857684	5
Neurotensin Receptor	12598608	5
NGF Receptor	9775979	5
P2 Purinoceptor	10717414	1
PGE2 Receptor	15234107	11
VEGF Receptor	12417438	4
Vitronectin Receptor	1705945	10
Schwann Cells		
5-HT Receptor	14724380	8
alpha1 Integrin	9187084	4
alpha1beta1 Integrin	9407013	1
alpha5 Integrin	9718369	8
alphav Integrin	9187084	4
Angiotensin II Receptor	7823177	1
ATP Receptor	9135063	19

continued on next page

Receptor name	PMID	Sentence
beta1 Integrin	9187084	4
c-met Protein	7996175	4
EGF Receptor	12612091	3
Endothelin Receptor	9130251	1
Glutamate Receptor	10535694	7
low affinity NGF Receptor	1377231	8
PDGF Receptor	8432400	2
trkB Receptor	8389459	9
VEGF Receptor	10742147	7

Bibliography

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev Mod Phys*, **74**, 47–97.

Albert, S., Gaudan, S., Knigge, H., Raetsch, A., Delgado, A., Huhse, B., Albers, H. K. M., Rebholz-Schuhmann, D., and Koegl, M. (2003). Computer-assisted generation of a protein-interaction database for nuclear receptors. *Mol Endocrinol*, **17**, 1555–1567.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Publishing, Inc., New York & London, 4th edition.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, **25**(1), 25–9.

Augustin, M., P.Heimann, Rathke, S., and Jockusch, H. (1997). Spinal muscular atrophy gene wobbler of the mouse: evidence from the chimeric spinal cord and testis for cell-autonomous function. *Dev Dyn*, **209**, 286–295.

Bachrach, C. and Charen, T. (1978). Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Med Inform (Lond)*, **3**.

Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., and Hogue, C. (2001). BIND–the biomolecular interaction network database. *Nucleic Acids Res*, **29**, 242–245.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barabási, A.-L. and Oltvai, Z. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, **5**, 101–113.

Bard, J., Rhee, S., and Ashburner, M. (2005). An ontology for cell types. *Genome Biology*, **6**.

Bassingthwaighte, J. (1995). Toward modeling the human physionome. In S. Sideman and R. Beyar, editors, *Molecular and Subcellular Cardiology: Effects on Structure and Function*, volume 382 of *Adv Exp Med Biol*, New York. Plenum Press.

Batagelj, V. and Mrvar, A. (2003). Pajek – analysis and visualization of large networks. In M. Jünger and P. Mutzel, editors, *Graph Drawing Software*, pages 77–103, Berlin. Springer.

Baumbach, J. (2005). *Graph based analysis of biological networks in the context of experimental results*. Master's thesis, Bioinformatics / Medical Informatics Department at Bielefeld University, Germany and BAB division at Rothamsted Research, UK.

Beal, M. F., Lang, A., and Ludolph, A., editors (2005). *Neurodegenerative Diseases: Neurobiology, Pathogenesis and Therapeutics*. Cambridge University Press.

Becker, K., Hosack, D., Dennis, G., Lempicki, R., Bright, T., Cheadle, C., and Engel, J. (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.

Blaschke, C. and Valencia, A. (2001). The Potential Use of SUISEKI as a Protein Interaction Discovery Tool. In *Proc 12th Genome Inform Workshop*, pages 123–134. Universal Academy Press, Tokyo, Japan.

Blaschke, C., Andrade, M., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. In *Pac Symp Biocomput*, pages 60–67.

Blaschke, C., Hirschman, L., and Valencia, A. (2002). Information extraction in molecular biology. *Briefings in Bioinformatics*, **3**, 154–165.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, **32**, D267–D270.

Boillée, S., Peschanski, M., and Junier, M.-P. (2003). The Wobbler Mouse. *Mol Neurobiol*, **28**, 65–106.

Bower, J. and Bolouri, H. (2000). *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, MA.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proc 6th Applied Nat Lang Proces Conference*.

Brooks, B., Feussner, G., and Lust, W. (1983). Spinal cord metabolic changes in murine retrovirus-induced motor neuron disease. *Brain Res Bull*, **11**(6), 681–686.

Chen, H. and Sharp, B. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, **8**(5), 147–160.

Chen, L., Liu, H., and Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*.

Christopher, R., Dhiman, A., Fox, J., Gendelman, R., Haberitcher, T., Kagle, D., Spizz, G., Khalil, I., and Hill, C. (2004). Data-driven computer simulation of human cancer cell. *Ann N Y Acad Sci*, **1020**, 132–153.

Cohen, A. and Hersh, W. (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, **6**(1), 57–71.

Collier, N., Nobata, C., and Tsujii, J. (2000). Extracting the names of genes and gene products with a hidden markov model. In *Proc COLING 2000*, pages 201–207, Saarbruecken, Germany.

Cooper, G. (2000). *The Cell – A Molecular Approach*. ASM Press, Washington D.C., USA, 2nd edition.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: an Architecture for Development of Robust HLT Applications. In *Proc 40th Anniv Meeting of the Assoc for Comp Ling*.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, **20**, 604–611.

Diambra, L. and da F. Costa, L. (2005). Complex networks approach to gene expression driven phenotype imaging. *Bioinformatics*, **21**(20), 3846–3851.

Diestel, R. (2000). *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer, 2nd edition.

Ding, J. and Berleant, D. (2005). MedKit: a helper toolkit for automatic mining of MEDLINE/PubMed citations. *Bioinformatics*, **21**, 694–695.

Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases? In *Pac Symp Biocomput*, pages 326–337.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G., Michalickova, K., Pawson, T., and Hogue, C. (2003). PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**(11).

Duchen, L. and Strich, S. (1968). An hereditary motor neurone disease with progressive denervation of muscle in the mouse: the mutant wobbler. *J Neurol Neurosurg Psychiat*, **31**, 535–542.

Eijk, C. V. D., Mulligen, E. V., Kors, J., Mons, B., and Berg, J. V. D. (2004). Constructing an associative concept space for literature-based discovery. *J Am Soc Inf Sci*, **55**, 436–444.

Falconer, D. (1956). Private communication. *Mouse News Lett*, **15**, 23.

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. Language, Speech and Communication. MIT Press, Cambridge, MA.

Festoff, B., D'Andrea, M., Citron, B., Salcedo, R., Smirnova, I., and Andrade-Gordon, P. (2000). Motor neuron cell death in wobbler mutant mice follows overexpression of the G-protein-coupled, protease-activated receptor for thrombin. *Mol Med*, **6**(5), 410–429.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(Suppl 1), S74–S82.

Gaizauskas, R., Demetriou, G., Artymiuk, P., and Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, **19**, 135–143.

Gasteiger, E., Jung, E., and Bairoch, A. (2001). SWISS-PROT: Connecting biological knowledge via a protein database. *Curr Issues Mol Biol*, **3**, 47–55.

Gilman, A., Simon, M., Bourne, H., Harris, B., Long, R., and Ross, E. (2002). Overview of the Alliance for Cellular Signaling. *Nature*, **420**, 703–716.

Goethe, J. W. (1986 (1808)). *Faust I*. Reclam, Stuttgart.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, **43**, 907–928.

Hamosh, A., Scott, A., Amberger, J., Bocchini, C., and McKusick, V. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, **33**, D514–D517.

Hanahan, D. and Weinberg, R. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Hanisch, D., Fluck, J., Mevissen, H., and Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. In *Pac Symp Biocomput*, pages 403–414.

Hatzivassiloglou, V. and Weng, W. (2002). Learning anchor verbs for biological interaction patterns from published text articles. *Int J Med Inform*, **67**, 19–32.

Heimann, P., Laage, S., and Jockusch, H. (1991). Defect of sperm assembly in a neurological mutant of the mouse. *Differentiation*, **47**, 77–83.

Henrotin, Y., Zheng, S., Labasse, A., Deby, G., Crielaard, J., and Reginster, J. (2000). Modulation of human chondrocyte metabolism by recombinant human interferon. *Osteoarthritis Cartilage*, **8**(6), 474–82.

Hiesinger, P. and Hassan, B. (2005). Genetics in the age of systems biology. *Cell*, **123**(7), 1173–1174.

Hofmann, O. and Schomburg, D. (2005). Concept-based annotation of enzyme classes. *Bioinformatics*, **21**, 2059–2066.

Hsing, M., Bellenson, J., Shankey, C., and Cherkasov, A. (2004). Modeling of cell signaling pathways in macrophages by semantic networks. *BMC Bioinformatics*, **5**, 156–168.

Huang, M., Zhu, X., Hao, Y., Payan, D., Qu, K., and Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, **20**, 3604–3612.

Hunter, P., Smith, N., Fernandez, J., and Tawhai, M. (2005). Integration from proteins to organs: the IUPS Physiome Project. *Mech. Ageing Dev*, **126**(1), 187–192.

Igarashi, T. and Kaminuma, T. (1997). Development of a cell signaling networks database. In *Pac Symp Biocomput*, volume 2, pages 187 – 197.

Jelier, R., Jenster, G., Dorssers, L., van der Eijk, C., van Mulligen, E., Mons, B., and Kors, J. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, **21**, 2049–2058.

Jenssen, T., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**(1), 21–28.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kaiser, M. and Hilgetag, C. (2004). Modelling the development of cortical system networks. *Neurocomputing*, **58**, 297–302.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The kegg resources for deciphering the genome. *Nucleic Acids Res*, **32**, D277–D280.

Karopka, T., Scheel, T., Bansemer, S., and Glass, A. (2004). Automatic construction of gene relation networks using text mining and gene expression data. *Med Inform Internet Med*, **29**, 169–183.

Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proc Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8, Philadelphia, USA.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien-Kruger, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, V., McCarthy, M., Hide, T., and Hide, W. (2003). eVOC: A Controlled Vocabulary for Gene Expression Data. *Genome Res*, **13**, 1222–1230.

Kitano, H. (2002). Computational systems biology. *Nature*, **420**(6912), 206–210.

Köhler, J., Philippi, S., and Lange, M. (2003). Semeda: ontology based semantic integration of biological databases. *Bioinfomatics*, **19**(19), 2420–2427.

Köhler, J., Rawlings, C., Verrier, P., Mitchell, R., Skusa, A., Rüegg, A., and Philippi, S. (2004). Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures. *In Silico Biology*, **5**, 005.

Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Rüegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**(11), 1383–1390.

Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *J Biomed Inform*, **37**(6), 512–526.

Krauthammer, M., Kra, P., Iossifov, I., Gomez, S., Hripcsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. (2002). Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, **18**(Suppl 1), S249–S257.

Leestma, J. and Sepsenwol, S. (1980). Sperm tail asoneme alterations in the wobbler mouse. *J Reprod Fert*, **58**, 267–270.

Leroy, G., Chen, M., and Martinez, J. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*, **36**, 145–158.

Lin, S., McConnell, P., Johnson, K., and Shoemaker, J. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, **20**, 3659–3661.

Ma, H.-W. and Zeng, A.-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **19**(11), 1423–1430.

Majoor-Krakauer, D., Willems, P., and Hofman, A. (2003). Genetic epidemiology of amyotrophic lateral sclerosis. *Clin Genet*, **63**, 83–101.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, USA.

Marcotte, E., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. *Bioinformatics*, **17**(4), 359–363.

McCray, A. and Nelson, S. (1995). The representation of meaning in the UMLS. *Methods Inf Med*, **34**, 193–201.

McDonald, D., Chen, H., Su, H., and Marshall, B. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, **20**, 3370–3378.

Meikle, A., de Sousa, J. C., Dacosta, N., Bishop, D., and Samlowski, W. (1992). Direct and indirect effects of murine interleukin-2, gamma interferon, and tumor necrosis factor on testosterone synthesis in mouse leydig cells. *J Androl*, **13**(5), 437–443.

Mika, S. and Rost, B. (2004). NLProt: extracting protein names and sequences from papers. *Nucleic Acids Res*, **32**, W634–W637.

Nenadic, G., Spasic, I., and Ananiadou, S. (2003). Terminology-driven mining of biomedical literature. *Bioinformatics*, **19**, 938–943.

Newman, M. (2003). The structure and function of complex networks. *SIAM REV*, **45**(2), 167–256.

Ng, S.-K. and Wong, M. (1999). Toward Routine Automatic Pathway Discovery from Online Scientific Text Abstracts. In *Proc 10th Genome Inform Workshop*, pages 123–134, Tokyo, Japan. Universal Academy Press.

Novichkova, S., Egorov, S., and Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts'. *Bioinformatics*, **19**, 1699–1706.

Noy, N., Musen, M., Mejino, J., and Rosse, C. (2004). Pushing the Envelope: Challenges in a Frame-Based Representation of Human Anatomy. *Data and Knowledge Engineering*.

Oliver, D., Bhalotia, G., Schwartz, A., Altman, R., and Hearst, M. (2004). Tools for loading MEDLINE into a local relational database. *BMC Bioinformatics*, **5**, 146.

Ono, T., Hishigaki, H., and Takagi, A. T. T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Oppenheim, J., Feldmann, M., Durum, S., Hirano, T., Vilcek, J., and Nicola, N., editors (2000). *The online Cytokine Reference Database*. Academic Press.

Papin, J. and Subramaniam, S. (2004). Bioinformatics and cellular signaling. *Cur Op Biotech*, **15**, 78–81.

Papin, J., Hunter, T., Palsson, B., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol*, **6**(2), 99–111.

Park, J., Kim, H., and Kim, J. (2001). Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Pac Symp Biocomput*, pages 396–407.

Peri, S., Navarro, J., Amanchy, R., Kristiansen, T., Jonnalagadda, C., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H., Rashmi, B., Ramya, M., Zhao, Z., Chandrika, K., Padma, N., Harsha, H., Yatish, A., and Kavitha, M. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**, 2363–2371.

Plake, C., Hakenberg, J., and Leser, U. (2005). Learning Patterns for Information Extraction from Free Text. In *Proc AKKD 2005*, Karlsruhe, Germany.

Rindflesch, T., Hunter, L., and Aronson, A. (1999). Mining molecular binding terminology. In *Proc 1999 AMIA Annual Symp*, pages 127–131, Bethesda, MD.

Ruch, P., Baud, R., and Geissbuhler, A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artif Intell Med*, **29**, 169–184.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P., Weng, W., Wilbur, W., Hatzivassiloglou, V., and Friedman, C. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform*, **37**, 43–53.

Santos, C., Eggl, D., and States, D. (2005). Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics*, **21**, 1653–1658.

Schacherer, F., Choi, C., Gotze, U., Krull, M., Pistor, S., and Wingender, E. (2001). The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, **17**(11), 1053–1057.

Schmitt-John, T., Drepper, C., Mussmann, A., Hahn, P., Kuhlmann, M., Thiel, C., M, M. H., Lengeling, A., Heimann, P., Jones, J., Meisler, M., and Jockusch, H. (2005). Mutation of vps54 causes motor neuron disease and defective spermiogenesis in the wobbler mouse. *Nature Genetics*, **37**(11), 1213–1215.

Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., and Schomburg, D. (2002). BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, **27**, 54–56.

Schuemie, M., Weeber, M., Schijvenaars, B., van Mulligen, E., van der Eijk, C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, **20**(16), 2597–2604.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol.*, **18**(12), 1257–1261.

Sekimizu, T., Park, H., and Tsujii, J. (1998). Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. In *Genome Inform Ser Workshop Genome Inform*, volume 9, pages 62–71.

Shatkay, H. and Feldman, R. (2003). Mining the biomedical literature in the genomic era: an overview. *J Comput Biol*, **10**, 821–855.

Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C. (2003). Effective adapatation of hidden markov model-based named entity recognizer for biomedical domain. In *Proc Workshop on Natural Language Processing in the Biomedical Domain*, pages 49–56, Saporo, Japan.

Shi, L. and Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics*, **6**, 88.

Sivakumaran, S., Hariharaputran, S., Mishra, J., and Bhalla, U. (2003). The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics*, **19**(3), 408–415.

Skusa, A., Köhler, J., and Rüegg, A. (2005). Extraction of biological interaction networks from scientific literature. *Brief Bioinform*, **6**(3), 263–276.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biol*. (accepted).

Srinivasan, P. and Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, **20**(Suppl 1), I290–I296.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34–45.

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., and Mostafa, J. (2001). Detecting gene relations from Medline abstracts. In *Pac Symp Biocomput*, pages 483–495.

Stone, V., Fishman, D., and Frese, D. (1998). Searching online and web-based resources for information on natural products used as drugs. *Bull Med Libr Assoc*, **86**, 523–527.

Suber, P. (2002). Open access to the scientific journal literature. *J Biol*, **1**(1), 3.

Swanson, D. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, **30**, 7–18.

Tan, A. (1999). Text mining: the state of the art and the challenges. In *Proc of the Pacic Asia Conf on Knowledge Discovery and Data Mining PAKDD 1999*, pages 65–70.

Taubert, J. (2005). *Database Integration and Analysis of Biological Networks Methods and Optimisation of ONDEX*. Master's thesis, Bioinformatics / Medical Informatics Department at Bielefeld University, Germany and BAB division at Rothamsted Research, UK.

Tieri, P., Valensin, S., Latora, V., Castellani, C., Marchiori, M., Remondini, D., and Franceschi, C. (2005). Quantifying the relevance of different mediators in the human immune cell network. *Bioinformatics*, **21**(8), 1639–1643.

Vintar, S., Buitelaar, P., and Volk, M. (2003). Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. In *Proc ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*.

Walsh, S., Anderson, M., and Cartinhour, S. (1998). Acedb: a database for genome information. *Methods Biochem Analysis*, **39**, 299–318.

Watts, D. and Strogatz, S. (1998). Collective dynamics of "small-world" networks. *Nature*, **393**, 440.

Weeber, M., Kors, J., and Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform*, **6**(3).

Werner, T. (2005). The next generation of literature analysis: Integration of genomic analyses into text mining. *Brief Bioinform*, **6**(3).

Williams, A. (2002). Defining neurodegenerative diseases. *British Medical Journal*, **324**, 1465–1466.

Wingender, E. (2004). TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol*, **4**(1), 55–61.

Winnenburg, R. (2005). *Text mining methods for extracting biological networks from scientific literature on plants and fungi*. Master's thesis, Bioinformatics / Medical Informatics Department at Bielefeld University, Germany and BAB division at Rothamsted Research, UK.

Winnenburg, R., Baldwin, T., Urban, M., Rawlings, C., Köhler, J., and Hammond-Kosack, K. (2006). PHI-base: A new database for pathogen host interactions. *Nucleic Acids Res*, **34**. Database issue, in press.

Wong, L. (2001). PIES, a protein interaction extraction system. In *Pac Symp Biocomput*, pages 520–531.

Wren, J., Bekeredjian, R., Stewart, J., Shohet, R., and Garner, H. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**(3), 389–398.

Wren, J., Chang, J., Pustejovsky, J., Adar, E., Garner, H., and Altman, R. (2005). Biomedical term mapping databases. *Nucleic Acids Res*, **33**, D289–D293.

Xenarios, I., amd X.J. Duan, L. S., Higney, P., Kim, S., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30**, 303–305.

Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. In *Pac Symp Biocomput*, pages 408–419.

Yao, D., Li, M., Lu, Y., Lin, N., Noble, N., Payan, D., Qu, K., Sun, H., Wang, J., and Zhu, X. (2004). PathwayFinder: Paving the Way Towards Automatic Pathway Extraction. In *Proc 2nd Asia-Pacific Bioinformatics Conference (APBC2004)*, volume 29, pages 53–62, Dunedin, New Zealand.

Yeh, A., Hirschman, L., and Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19**(Suppl 1), i331–i339.

Yu, H. and Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, **19**(Suppl 1), i340–i349.

Zankl, M., Petoussi-Henss, N., Fill, U., and Regulla, D. (2003). The application of voxel phantoms to the internal dosimetry of radionuclides. *Radiation Protection Dosimetry*, **105**, 539–448.