0888-8809/03/$15.00/0
Printed in U.S.A.

Molecular Endocrinology 17(8):1555–1567

doi: 10.1210/me.2002-0424

Computer-Assisted Generation of a Protein-Interaction Database for Nuclear Receptors

SYLVIE ALBERT, SYLVAIN GAUDAN, HEIDRUN KNIGGE, ANDREAS RAETSCH, ASUNCION DELGADO, BETTINA HUHSE, HARALD KIRSCH, MICHAEL ALBERS, DIETRICH REBHOLZ-SCHUHMANN, AND MANFRED KOEGL

LION Bioscience AG, 69123 Heidelberg, Germany

With the increasing amount of biological data available, automated methods for information retrieval become necessary. We employed computer-assisted text mining to retrieve all protein-protein interactions for nuclear receptors from MEDLINE in a systematic way. A dictionary of protein names and of terms denoting interactions was generated, and trioccurrences of two protein names and one interaction term in one sentence were retrieved. Abstracts containing at least one such trioccurrence were manually checked by biologists to select the relevant interactions out of the automatically extracted data.

In total, 4360 abstracts were retrieved containing data on protein interactions for nuclear receptors. The resulting database contains all reported protein interactions involving nuclear receptors from 1966 to September 2001. Remarkably, the annual increase in number of reported interactors for nuclear receptors has been following an exponential growth curve in the years 1991 to 2001.

Apparent in the data set is the high complexity of protein interactions for nuclear receptors. The number of interactions correlates with the number of published papers for a given receptor, suggesting that the number of reported interactors is a reflection of the intensity of research dedicated to a given receptor. Indeed, comparison of the retrieved data to a systematic yeast two-hybrid-based interaction analysis suggests that most NRs are similar with respect to the number of interacting proteins. The data set obtained serves as a source for information on NR interactions, as well as a reference data set for the improvement of advanced text-mining methods. (*Molecular Endocrinology* 17: 1555–1567, 2003)

NUCLEAR RECEPTORS (NRs) are ligand-inducible transcription factors that regulate the expression of target genes involved in a wide range of processes such as metabolism, development, reproduction, and cell differentiation. The activity of many of these pharmaceutically important transcription factors is regulated by the binding of small molecules to their ligand-binding domain (1). The altered conformation resulting from ligand binding influences the interactions of NRs with other proteins, named cofactors (2, 3). These interactions are required for transcriptional regulation of NR-responsive genes. Cofactors include chromatin-modifying proteins, proteins directly interacting with the basal transcriptional machinery, proteins involved in cytoplasmic signal transduction, as well as several proteins with poorly described function.

The responsiveness of NRs to small molecule ligands makes them excellent drug targets, as exemplified by the many successful drugs that target nuclear receptors (4). Interestingly, different ligands for the same NR can have diverse biological effects. For example, the natural ligand estradiol activates the estrogen receptor (NR3A/ER) in the breast epithelium as well as in bone, whereas the synthetic ligand raloxifen represses ER function in the breast, but has agonistic effects in bone (5–7). The binding of the different ligands seems to induce different conformations of the receptor, which then result in an altered preference of the receptor for the available cofactors (8–12). These ligand-dependent cofactor preferences are part of a possible explanation for differences in the biological effects of different ligands on the same NR (13). Therefore, protein-protein interactions involving NRs and cofactors are of particular interest in drug discovery.

Their outstanding biological and pharmaceutical importance has made NRs, and particularly steroid hormone receptors, a very well researched group of proteins. Although there are a number of excellent NR-specific databases (14–16), there is no publicly available resource dealing with the cofactor specificity of each NR. Primary scientific literature is the best source of information in this case. A widely used literature resource is MEDLINE (>11 million abstracts accessible via PubMed, at http://www.ncbi.nlm.nih.gov/Entrez), which represents a vast corpus of medical and molecular biology literature available electronically. However, the sheer size of the data poses problems. For example, a mere query in MEDLINE using the term

Abbreviations: ER, Estrogen receptor; ERR, estrogen-related receptor; HNF, hepatocyte nuclear receptor; LXR, liver X receptor; MR, mineralocorticoid receptor; NCoR, nuclear receptor corepressor; NR, nuclear receptor; NRBP, NR-binding protein; PPAR, peroxisome proliferator-activated receptor; RAR, retinoic acid receptor; ROR, retinoid-related orphan receptor; SRC-1, steroid receptor coactivator 1; TR, thyroid hormone receptor; TRAP, TR-associated protein; Y2H, yeast two hybrid.

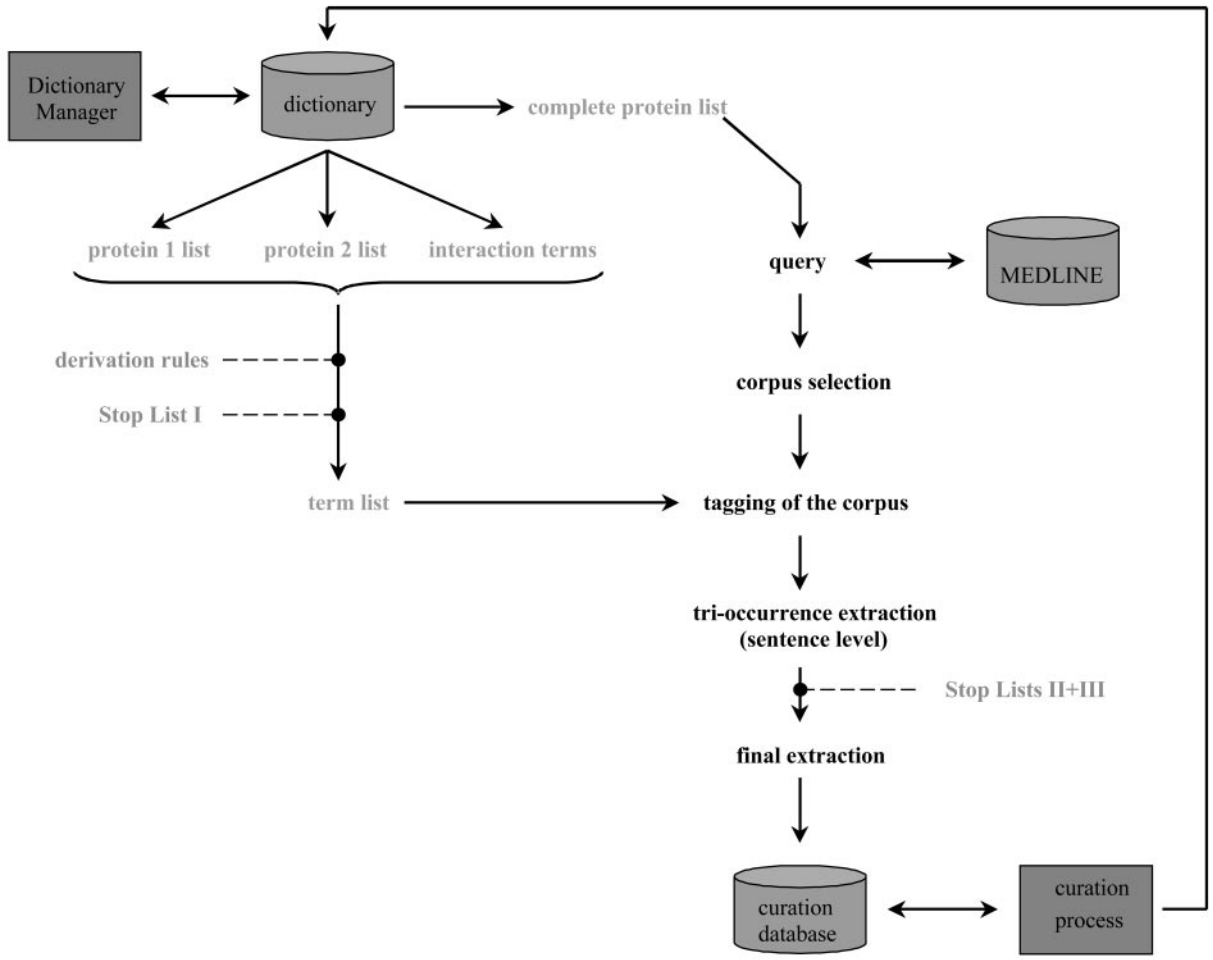

Fig. 1. Flowchart of the Extraction and Curation Process

The dictionary contains the terms described in *Materials and Methods*. The Dictionary Manager is the web interface used to access and to modify the dictionary data. The dictionary was used on one side to provide the complete protein name list used to query MEDLINE and on the other side to extract the term list of interest for the tagging of the corpus resulting from the query. Trioccurrences were extracted at the sentence level, some were eliminated by applying stop lists, and the resulting trioccurrences were dumped into the curation database and curated. During the curation process, relevant terms were found and added to the dictionary: *round box*, database; *rectangular box*, web interface; *solid circle with dashed line*, rules applied.

"estrogen receptor" yields more than 12,000 citations, and queries for other NRs as well return several thousand abstracts. The returned abstracts have to be subselected for the retrieval of papers covering only certain aspects of NR biology. However, there is no easy way, at present, to extract all published cofactors for a given NR, or to retrieve only the abstracts concerning NR-protein interactions, for example. This shortcoming is reflected in recent interest in the development of text-mining technologies (17–21). Successful applications of text mining for protein interactions have been reported (22, 23).

In an attempt to overcome the above mentioned difficulties, we applied automated text-mining methods for the retrieval of all abstracts reporting NR-cofactor interactions. The automatically retrieved data were quality controlled and completed by biologists. In the course of the project, a dictionary was created containing NRs, cofactors, and other NR-binding proteins (for simplicity, all proteins interacting with NRs including cofactors are referred to as "NRBPs" in this paper) and their synonyms, as well as expressions describing protein-protein interactions. The project yielded a database resource containing all NR-NRBP interactions for the protein names contained in the dictionary published in MEDLINE abstracts between 1966 and September 2001.

This paper describes our general approach, as well as the extraction and curation processes, and discusses the resulting protein interaction data.

RESULTS

The goal of this project was the generation of a knowledge database allowing scientists to get an overview

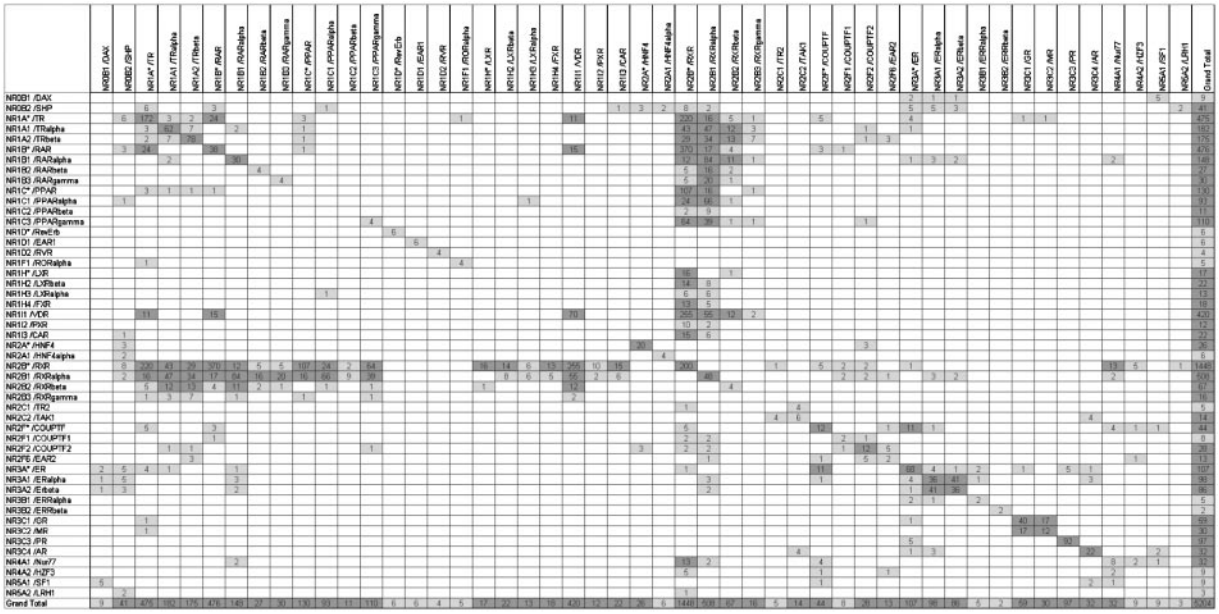

Fig. 2. NR-NR Interactions

Homo- and heterodimerizations within the NR family are plotted. Some of the entities refer to families (*e.g.* TR) and some refer to the gene locus/protein level (see *Materials and Methods*, *e.g*. TRα). *Numbers* indicate the number of times a particular interaction was mentioned in our data set. Shades of *gray* are according to these numbers: 1–10 = *pale gray*, 11–50 = *intermediate gray*, more than 50 = *dark gray*. The same table is available in the online supplemental data, with a direct reference to the PubMed identification number of the respective abstracts. Note that the table is redundant, because it is symmetrical across the diagonal.

on the publicly known information concerning protein-protein interactions in the NR domain. The technology chosen to extract the information from scientific abstracts in MEDLINE was the automated cooccurrence extraction of three entities within a sentence (trioccurrence extraction), namely two protein names and a verb/noun expressing an interaction. This trioccurrence extraction method enhances the recall (number of relevant documents retrieved per number of total relevant documents) at the expense of the precision (number of relevant documents retrieved per total number of retrieved documents). Subsequently, data were curated by biologists evaluating the accuracy of the data. Curation by biologists is necessary to increase the precision to a value close to 100% by distinguishing between sentences mentioning relevant interactions such as "*ARA70* which specifically *interacts* with *androgen receptor* was also cloned recently" (the three entities of interest are linked) and sentences mentioning nonrelevant interactions such as "*TRAP* is able to *bind* to DNA, even in the absence of functional *TR*" (the three entities of interest are not linked).

NR Dictionary

The trioccurrence approach requires the compilation of a list of terms to be searched for in the sentences. Better than a mere list of terms, it was decided to build a dictionary with a hierarchical structure allowing as much quality as possible concerning the description of the entities. We exclusively focused on the three most studied mammals for NRs, *i.e.* human, mouse, and rat. In case the species is not specified in the selected sentence, but from the context it is clear that it is a mammalian species, the generic term "mammal" is applied.

At the end of the project, the dictionary contained 563 terms for 49 NRs orthologs (plus 11,928 synonyms) and 570 NRBPs (plus 4,415 synonyms). For details on the dictionary, see *Materials and Methods*.

Extraction Process

The extraction process was run on all MEDLINE abstracts available at the beginning of the project, *i.e.* those found in the literature database from 1966 until September 10, 2001. The abstracts and sentences of interest were retrieved by applying the following sequence of operations (see *Materials and Methods*):

1) Selection of the abstracts containing at least one protein name from the dictionary (= corpus selection)

2) Tagging of the selected corpus using the protein names and the interaction terms from the dictionary

3) Trioccurrence (protein1 + protein2 + interaction term) extraction at the sentence level

A flow chart describing the complete process is presented in Fig. 1.

New entities were added into the dictionary during the curation process, resulting in an improved dictionary at the end of the curation. To make the extraction

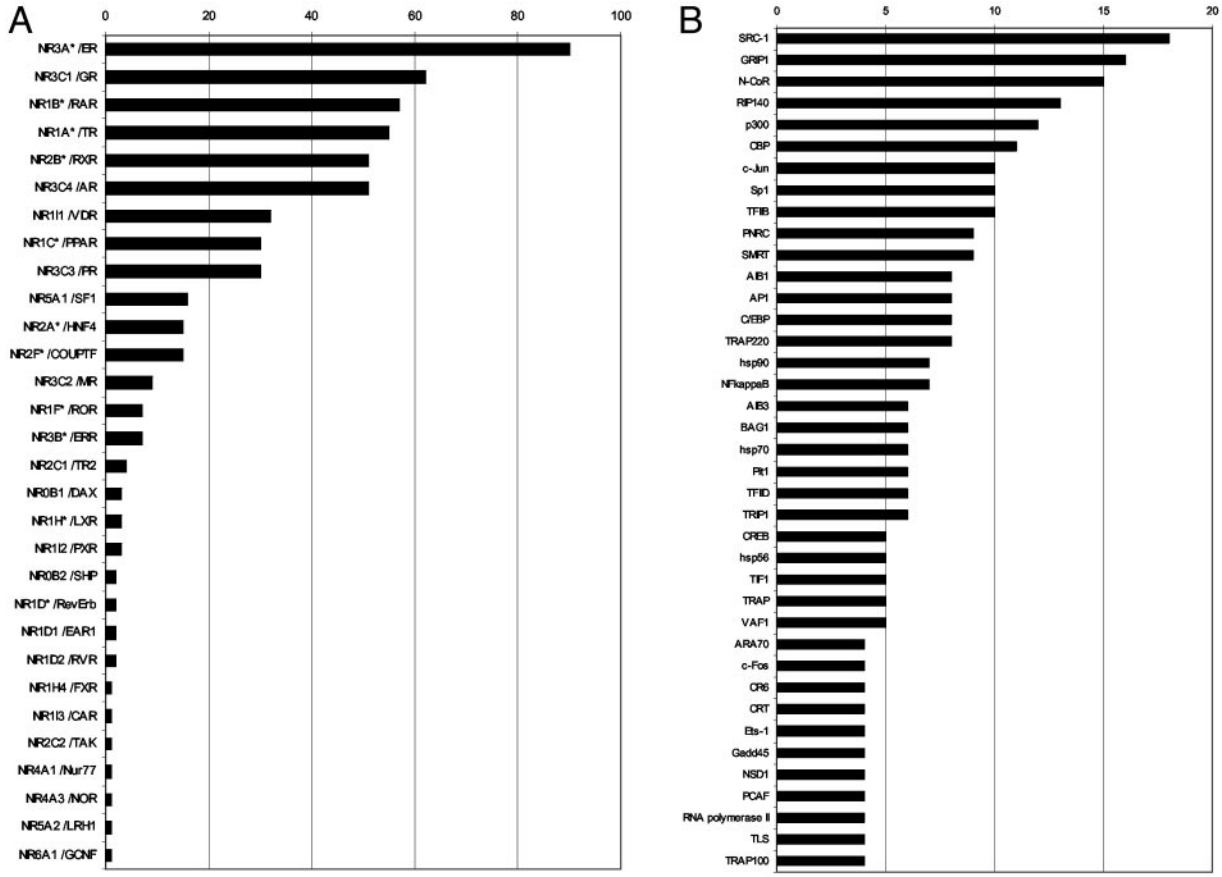

Fig. 3. NR-NRBP Interactions

A, Number of NRBPs published to bind to a NR (histographical representation of the number of NRBPs for each family of receptors). B, Number of NRs published to bind to a NRBP (histographical representation of the number of NR families per NRBP). In the case of the NRBPs, some entities refer to the family level or to complexes (*e.g.* TRAP), whereas others refer to entities at the gene locus/protein level (*e.g.* TRAP220; see *Materials and Methods* for the organization into description levels). For this representation, the following receptors were summarized as families: NR1A/TR (NR1A1 and NR1A2), NR1B/RAR (NR1B1, NR1B2, and NR1B3), NR1C/PPAR (NR1C1, NR1C2, and NR1C3), NR1F/ROR (NR1F1 and NR1F2), NR1H/LXR (NR1H2 and NR1H3), NR2A/HNF4 (NR2A1 and NR2A2), NR2B/retinoid X receptor (NR2B1, NR2B2, and NR2B3), NR2F/chicken ovalbumin upstream promoter transcription factor (NR2F1 and NR2F2), NR3A/ER (NR3A1 and NR3A2), and NR3B/ERR (NR3B1, NR3B2, and NR3B3).

of trioccurrences as complete as possible, we used the improved dictionary to reextract trioccurrences from the previously selected abstracts after the first round of extraction and curation. The newly extracted trioccurrences were quality controlled by manual curation. Because the new terms in the dictionary comprised only NRBPs but no novel NRs, we expect that most of the abstracts containing NR-NRBP interactions were analyzed. However, because the corpus extraction was not repeated with the new dictionary, we will likely miss abstracts describing NRBP-NRBP interactions due to the absence of the respective abstracts in the corpus.

Curation Process

The curation by domain experts is essential to select relevant interactions because the mere presence of three entities of interest in the same sentence does not guarantee that these entities describe a protein-protein interaction. The process leads to the production of a high-precision data set of protein-protein interactions in the NR domain that also represents a training set for further text-mining investigations. The curation additionally permits the evaluation of the automatic extraction technique.

Two biologists curated the whole set of automatically extracted abstracts independently of each other with the help of a customized graphic user interface optimized for their needs. A third biologist resolved the conflicts between the two curators and handled the dictionary's improvements and modifications in collaboration with them.

Curators also added novel interactions that had not been found by the automatic process but were present within the selected abstracts. This was, in some cases, due to the lack of a protein name in the dictionary, which was added at that time. In other cases, the three

terms necessary to express a protein-protein interaction were found in separate sentences, which explains why the automatic extraction program could not pick up these interactions. To get a consistent data set, curators also removed interactions referring to nonmammal proteins. Thus, there are no data on insect NRs, for example, in our curated data set.

Primary Results of the Extraction

A total number of 4,360 abstracts was retrieved and processed, containing 15,608 automatically extracted trioccurrences, which represents an average of 3.6 trioccurrences per abstract. After curation, 3308 trioccurrences were classified as showing a positive interaction (A binds to B) and 143 as denying a relation (A does not bind to B), corresponding to an overall precision of 22%. The curators were furthermore adding 3556 trioccurrences expressing a positive interaction and 163 expressing a negative interaction. The complete data set of validated interactions obtained after curation, *i.e.* 7170 interactions, is available as supplemental data, which are published on The Endocrine Society's Journals Online web site at http://mend.endojournals.org. Interactions that are denied in a paper are included as a supplemental table in this file.

The most frequently used terms to describe an interaction were "dimerize" and "interact." Together, they accounted for 57% of all trioccurrences. The most reliable term for the automated retrieval of interactions was "dimerize," whereas "link," "couple," and "affinity," led to a low percentage (<5%) of good extractions.

NR-NR and NRBP-NRBP Interactions

Figure 2 and supplemental table entitled "NR-NR-interactions" (see supplemental data) show the interactions of NRs with other NRs extracted from the abstracts. We have ordered NRs according to the official NR nomenclature in all Pivot tables (*e.g*. in Fig. 2). Because the official NR nomenclature reflects the phylogenetic relationship of NRs, these tables represent the protein interactions in an order based on phylogeny. For example, the tendency of the NR3 and NR1A families to form homodimers is clearly visible in the diagonal of the grid, as well as the heterodimerization of members of the NR1 family with NR2B/retinoid X receptor. Smaller groups of interaction are the heterodimers involving the NR1A/thyroid hormone receptor (TR), NR3A/ER, and NR2F/chicken ovalbumin upstream promoter transcription factor families. A similar plot is available for the interested reader in the online supplemental data for interactions within NRBPs. We believe that the digest of the literature presented here and below should be of interest as a reference especially for novices in this well-researched area of biology, complementing the excellent reviews that are available (*e.g*. Ref. 2).

NR-NRBP Interactions

Figure 3A and supplemental table entitled "NRs-NRBPs-interactions" (online supplemental data) show the number of different interacting proteins for each nuclear receptor family in our data set. The number of NRBPs per NR family ranges from 90 for NR3A/ER to 1 for NR6A1/germ cell nuclear factor or NR4A3/neuron-derived orphan receptor. For NR1B2/retinoic acid receptor-β (RARβ), NR1F3/retinoid-related orphan receptor (ROR)γ, NR2A2/hepatocyte nuclear factor (HNF)4γ, and NR2F6/EAR2 we did not find any reported interactions in the literature. The number of NR families published to interact with a given NRBP is plotted in Fig. 3B. Not surprisingly, the greatest number of receptors is found for the p160 family of coactivators [18 for steroid receptor coactivator 1 (SRC-1), 16 for glucocorticoid receptor-interacting protein 1/transcriptional intermediary factor 2] followed by 13 for receptor-interacting protein 140, 12 for p300, and

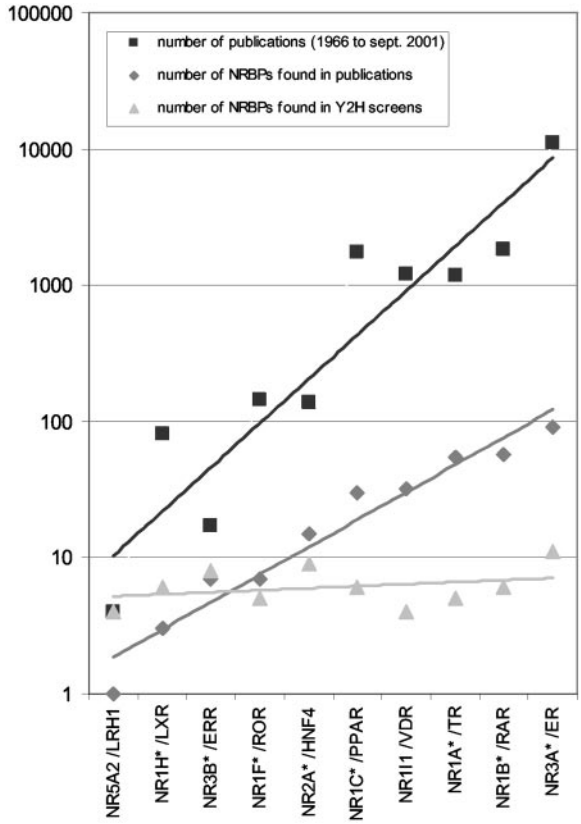

Fig. 4. Correlation of Known NRBPs with the Number of Publications for a Given NR

The *histogram* shows the number of papers published on a given receptor (*squares*), in comparison to the number of known NRBPs as present in our data set from text mining (*diamonds*) and the number of known NRBPs picked up in Y2H screens with these receptors (*triangles*). NRs are sorted according to the number of different proteins reported to bind to them. Details on the Y2H screens are found in *Materials and Methods* and in Table 1.

Table 1. Receptors, Libraries, and Ligands Used in Y2H Screens

NRs	Ligand	Libraries	SRC1	TIF2	NCoA3	RAP250	RIP140	PNRC	PNRC2	TRIP1
Ligand induced										
ER α-LBD	E_2	B, K, S, T	X						X	
ER β-LBD	E_2	H, L, O, T	X	X		X	X	X	X	X
LXR α-LBD	Tul	S, K, O, T	X				X			X
LXR β-LBD	HCh	B, F, O, S	X		X					
VDR-LBD	$VitD_3$	F, K, L, T	X				X			
TRα-LBD	T_3	F, L, O, T								
TRβ-LBD	T_3	F, L, O, T	X		X					X
PPARb-LBD	LA	K, L, O, T	X			X				X
RARe-LBD	RA	B, H, K, S								X
RARg	RA	B, H, O, T	X					X	X	X
Constitutive										
LRH1-LBD	None	B, H, K, L	X				X	X	X	
HNF4a-LBD	None	C, H, K, L				X	X	X	X	
HNF4g	None	L, O, S, T	X			X	X	X	X	X
ERRa-LBD	None	B, K, L, T	X							X
ERRg-LBD	None	C, F, K, T	X				X	X	X	X
RORa-*LBC	None	B, F, O, T,	X					X	X	X
RORb-LBD	None	F, K, O, T					X	X	X	X

Interactions of a receptor with a given cofactor are indicated by X. The total numbers of previously known NRBPs identified in the indicated libraries by Y2H screens are summed up to the *right,* both for the individual receptor and for the subfamily. For example, five NRBPs have been picked up for NR1A1/ERα, eight for NR1A2/ERβ, and 11 for either of the two. NR0B1/DAX1 and NR0B2/SHP have been included as NRBPs in this list since they are known to bind to NR LBDs in a mode related to that of classical cofactors.

Ligands and concentrations used in the screens are: E_2 (259 nM); Tularik, Tularik 901317 (1 μM); HChol, 22 *S*-hydroxy cholesterol (10 μM); VitD3, vitamin D_3 (250 nM); LA, linoleic acid (10 μM); RA, all-*trans*-retinoic acid (1 μM). Libraries screened are: B, brain; C, chondrocytes; F, fetal brain; H, heart; K, kidney; L, liver; O, ovary; S, skeletal muscle; T, testis; LBD, ligand-binding domain. Details on which interacting protein was identified in which library can be found in the online supplemental material, file "all interactions.xls" table "Y2H details."

11 for CREB-binding protein. Among the corepressors, nuclear receptor corepressor (NCoR) and silencing mediator of retinoid and thyroid hormone receptor seem to be promiscuous as well, because both proteins have been published to interact with 15 and nine different entities, respectively. The interactions mentioned most frequently are the interactions of heat shock protein 90 with NR3C1/glucocorticoid receptor and NR3C3/progesterone receptor (361 times and 86 times), the interactions of NR1A/TR with the TR-associated protein (TRAP) family of proteins, which are mentioned 78 times, the interactions of NR3A/ER with SRC-1 (59 times), and the interactions of NR1A/TR with NCoR, which are mentioned 60 times. Note that Fig. 3 contains entities that refer to the gene (locus) level, such as NR0B1/DAX1 and TRAP220, and entities that refer, respectively, to the family level or to complexes such as NR3A/ER and TRAP (see *Materials and Methods* for the organization into description levels). This is due to the fact that we could not distinguish which members of the family or complex were addressed in the respective abstracts. For consistency, NRs were grouped into families for the representation in Fig. 3.

As can be seen from Fig. 3A, a much higher number of interacting proteins is found for the classical hormone and metabolite receptors, such as NR1A/TRs, NR3A/ERs, and NR1C/peroxisome proliferator-activated receptors (PPARs) than for other receptors. A notable exception is NR3C2/mineralocorticoid receptor (MR), which is a well studied protein, but has as few reported interactions as the orphan receptor families NR3B/estrogen-related receptor (ERR) and NR1F/ROR. Also our in-house yeast two-hybrid (Y2H) studies hint toward a rather specific nature of MR. We screened four cDNA libraries for interactors of MR and isolated a total of 67 clones of which 59 encoded for fragments of only three known NRBPs, namely 53 times SRC-1, three times forkhead homolog in rhabdomysarcoma, and three times thyroid hormone receptor-interacting protein-1 (Koegl, M., and M. Albers, unpublished). The high number of interacting proteins for some of the receptors prompts the question whether these receptors have an extraordinary complex repertoire of NRBPs compared with the others, or whether the difference in the number of interacting proteins is just a reflection of the intensity of research that has been devoted to the different receptors. If the second hypothesis is correct, one would expect a correlation of the number of papers published on a given receptor with the number of interacting proteins. As can be seen in Fig. 4, such a correlation is evident from such a plot. This might suggest that for the less well studied receptors, an equally high number of interacting proteins can be expected to be discovered.

Table 1. Continued

PGC1	FKHR	FKHRL1	NRBF2	SHP	DAX1	DUT	CAMK2B	NCoR	NRBPs per receptor	NRBPs per NR Family
			X			X	X		5	11
X									8	
		X						X	5	6
		X							3	
				X					4	4
								X	1	5
							X		4	
	X		X	X					6	8
								X	2	6
				X					5	
									4	4
X				X					6	9
X	X			X					9	
				X		X			4	8
					X	X			7	
									4	5
									4	

To test this assumption in a more direct fashion, we looked for an independent method to predict the number of NRBPs for each receptor. In our laboratory, we have undertaken a systematic study of NR-NRBP interactions by Y2H screens (Koegl, M., and M. Albers, unpublished data). We wanted to compare the data produced by the systematic Y2H screening approach to the published data. However, for many less well studied receptors, no ligands are known, and the absence of a ligand in the Y2H screen might preclude the detection of many interactions, resulting in a bias. To prevent this bias, we concentrated on receptors that either have a known ligand that could be used in a screen [NR3A/ERs, NR1H2/liver X receptor (LXR)β, NR1H3/LXRα, NR1I1/vitamin D receptor, NR1A/TRs, NR1C/PPARs, NR1B/RARs], or have a high constitutive NRBP-binding activity (NR5A2/liver receptor homolog, NR2A/HNF, NR3B/ERRs, NR1F/RORs). Details and results of the screens are summarized in Table 1 (see *Materials and Methods*). As expected, we reproduced many of the published interactions with known NRBPs in our screens. These known NRBPs were also picked up a number of times as interactors of NRs to which they were not previously known to bind (Table 1). We then compared the number of interactions seen in Y2H screens to the number of published interactions (Fig. 4). As can be seen, the number of known NRBPs identified in these screens is approximately equally distributed across the graph, showing no correlation to the number of papers published for a given receptor. Thus, we may expect that as more research is devoted to the less well-characterized receptors, an equally complex pattern of interacting proteins will emerge as is already known for well studied receptors. This is also corroborated by considering the number of novel NRBPs reported per year in the time from 1987 to 2001. Figure 5 shows that the increase in reported NRBPs has been following an exponential curve in the years 1990 through September 2001. Although it is not possible to extrapolate such a curve in a meaningful manner, it is reasonable to predict that the number of reported NR-NRBP interactions is likely to increase substantially in the coming years, adding to the amazing complexity of protein interactions in this field of biology.

DISCUSSION

The systematic storage of DNA and protein sequence data in dedicated databases was crucial for the development and application of advanced bioinformatics. Presently, systematic approaches are being initiated to integrate sequence-based information with other data, such as expression, modification, or interaction data, often referred to as the emerging field of "systems biology" (24–26). Such attempts suffer greatly from the lack of systematic databases on biological knowledge. Scientific literature, as present in the abstracts in MEDLINE, is still the richest source of data in most areas of biology. In the present paper, we have applied computerized text mining to try and overcome some of the limits in data retrieval. The resulting data set may be used as a reference for the published interactions of NRs and their NRBPs, and guide users to relevant publications. We believe that this may be especially interesting for novices to the field of NRs when faced with the tremendous amount of literature published. The importance of NR-NRBP interactions in drug development should make this resource even more useful.

Limits of the Text-Mining Method

The extraction method applied ensures a high recall, at the expense of the precision, which reached a level of 22% before curation. Although promising, for many

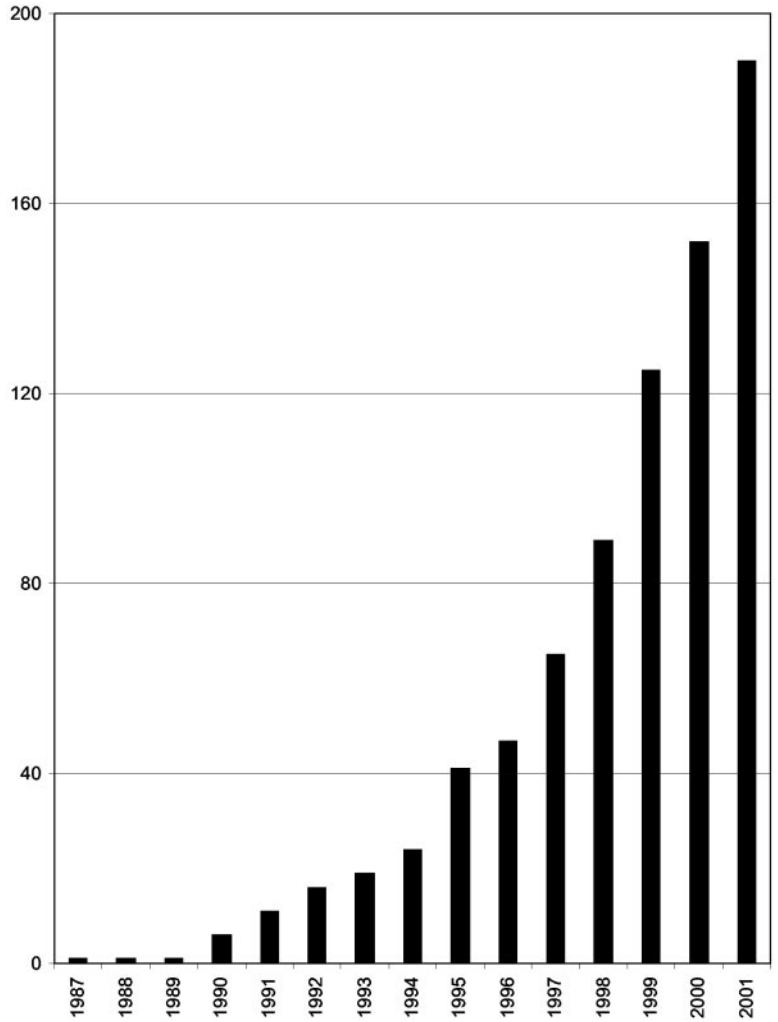

Fig. 5. Annual Increase of Reported NRBPs

The number of known NRBPs was calculated for our data set for each year, *i.e*. the annual increase in NRBPs reflects the discovery of interactions of NRs with proteins previously not known to bind to NRs. Note that the year 2001 covers only the period from January to the 10th of September.

applications of the data a precision of 22% is too low. More sophisticated methods are presently becoming available that should allow the precision to improve. However, to test such methods, a reference set of data is needed to determine the precision and recall of the method in order to allow assessment of potential improvements in the procedure. With the generation of the quality-controlled set of data presented here, we provide such a reference, which is, to our knowledge, the first such resource.

The presence of a high number of trioccurrences picked up that do not denote protein-protein interactions in the sentence is mainly due to the fact that the applied method does not analyze the sentence structure, but rather extracts all the possible triplet combinations. One of the frequently found problems is the coordination problem, exemplified by the following sentence: "A binds to B and C." The extraction program will provide the curators with two correct trioccurrences, "A, B, bind" and "A, C, bind", but also with "B, C, bind," which is obviously not correct. Another problem is the presence of more than three entities from the dictionary in the same sentence, but spread over several phrases, leading to the extraction of a high number of false trioccurrences.

Technologies able to extract information from text more sophisticated than the one presented in this paper are becoming available at present. Information Extraction is one of them (18, 19). Information Extraction takes advantage of Natural Language Processing techniques to produce a structured representation of pieces of free text. The input text is syntactically and semantically analyzed to locate the entities of interest. This approach is expected to result in a higher precision, but most likely at the expense of the recall (Kirsch, H., and S. Albert, unpublished data). In another scenario, automatic clustering of documents allows the user to have a good overview on a large

collection of documents and to make use of content words and content similarities between documents. In any case, the imprecise use of the scientific language in abstracts will put an implicit limit to even the most advanced methods of text mining. Reference to gene/protein families instead of precise identification of genes/proteins including variants, lack of species information, and unresolved or unused nomenclatures will not permit the extraction of precise data from abstracts.

The Complexity of Protein Interaction Networks

From our analysis the complexity of protein interactions of NRs is evident, and it appears that the number of reported interactions is likely to increase. This is, to a great extent, a reflection of the coming of age of high-throughput methods to detect protein-protein interactions, mostly the Y2H system, first published by Fields and Song in 1989 (27), but also recent improvements in mass spectrometry-based methods (28, 29). Thus, this surge in reported protein interaction data driven by proteomic methods parallels the increase in DNA sequence data generated by advanced DNA-sequencing technology. In contrast to DNA sequence information, however, systematic and comprehensive databases for protein interactions are only beginning to emerge (30, 31).

The Need for Systematic Databases

At present, automated methods can only deliver data with a limited precision. Some reliability can be gained for well researched interactions, *e.g*. by scoring the number of publications on a given interaction. For example, the automatic extraction of the interaction of NR1A/TR with NCoR is plausible, because it has been found in 60 different statements. Even though truth may not depend on the number of times a fact is stated, scientific consensus usually implicates reliability, exceptions notwithstanding. For less well researched data that are not mentioned several times in the scientific literature, methods as the one presented here can merely guide a scientist to the appropriate literature. We believe that this is of use, especially in combination with meaningful clustering methods of abstracts. However, to arrive at the creation of complete and reliable databases, *e.g*. on protein interactions, we believe that data will have to be entered manually. In a preferred setting, newly discovered data on protein-protein interactions will have to be deposited at a central resource at the time they are discovered and published, as has become good practice for newly discovered nucleotide sequences.

MATERIALS AND METHODS

Dictionary

The NR dictionary contained terms denoting proteins related to the NR domain (NRs and NRBPs) as well as terms expressing an interaction and terms denoting species of interest. Each term is associated with a synonym list. The synonym lists include synonyms (*e.g.* TRβ, c-erbAβ, and NR1A2) and orthographical variants (e.g. TRβ, TR β, and TR-β). New terms and synonyms were added during the curation process. The terms expressing an interaction comprise 13 verbs: to interact, to bind, to link, to contact, to couple, to assemble, to attach, to complex, to dimerize, to associate, to dock, to precipitate, and to dissociate. These verbs are conjugated in all tenses. The nominal forms deriving from these verbs are also considered as their synonyms. For instance, the verbs "to bind" and "to associate" exhibit the following forms as synonyms: "bind," "binds," "bound," "binding" and "associate," "associates," "associated," "associating," "association," respectively. The term "affinity" was initially also used as an interaction term, but removed from the list later, because it yielded too many false results. The terms "yeast two hybrid screening," "mammalian two hybrid screening" or "fluorescence resonance energy transfer (FRET)" were used when the sentence was expressing an interaction without containing any of the previously cited verbs/nouns but containing one of these method names instead. An example is the following sentence: "Using this region of PPARγ as bait, we have used a yeast two-hybrid screen to clone a novel protein, termed PGC-2, containing a partial SCAN domain." Each term denoting a protein is assigned to one species. Unfortunately, the precise denotation of proteins is often not possible from MEDLINE abstracts. For example, the description of interactions involving "estrogen receptor" (ER) is ambiguous because "estrogen receptor" is a generic term grouping several entities: there are two genes that could be referred to, ERα and ERβ. In addition, there are ERβ1, ERβ2, . . . and ERβ5, which are splice variants produced from the ERβ locus. Because, in general, authors of scientific abstracts do not assign a high resolution to the terms they are using, we decided to link each biological entity in the dictionary to one of the three following classes (Fig. 6): family level (ER), gene locus/protein level (ERα and ERβ), or gene variant level (ERβ1 to ERβ5), to be able to extract the correct information from the abstracts and to interpret and discuss the results in a suitable way. In addition, the following relations between the above cited entities were considered: 1) "kind-of" or "is-a" relation: one entity is an example of another entity ("human androgen receptor" is a "androgen receptor," "androgen receptor" is a "nuclear receptor"); 2) "part-of" relation: one entity is a part of another entity ("human androgen receptor" is a part of "human," "NFκB subunit p50" is a part of "NFκB"). This second relation is used to relate proteins to species in which they exist and to complexes.

The dictionary was populated based on literature knowledge. It was maintained and updated throughout the project in close collaboration with the curators. It was set up using an XML-based (eXtensible Markup Language) format. It can be accessed and modified via a graphic user interface that allows the user to search and modify it in a convenient way.

Extraction of the Trioccurrences

MEDLINE abstracts were downloaded according to the rules described at http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html and indexed under SRS (Sequence Retrieval System, http://srs.ebi.ac.uk).

During the trioccurrence extraction, all the NRs and NRBPs (= protein 1) were searched for in the sentences in combination with all the proteins entered in the dictionary (= protein 2: NRs, NRBPs, NR/NR complexes, NR/NRBP complexes, NRBP/NRBP complexes), and the 13 families of verbs/nouns and method names described above were used to pinpoint an interaction.

Some word-derivation rules necessary for the extraction were included in the corpus extraction program (see Fig. 1). They are of three types and allow the following:

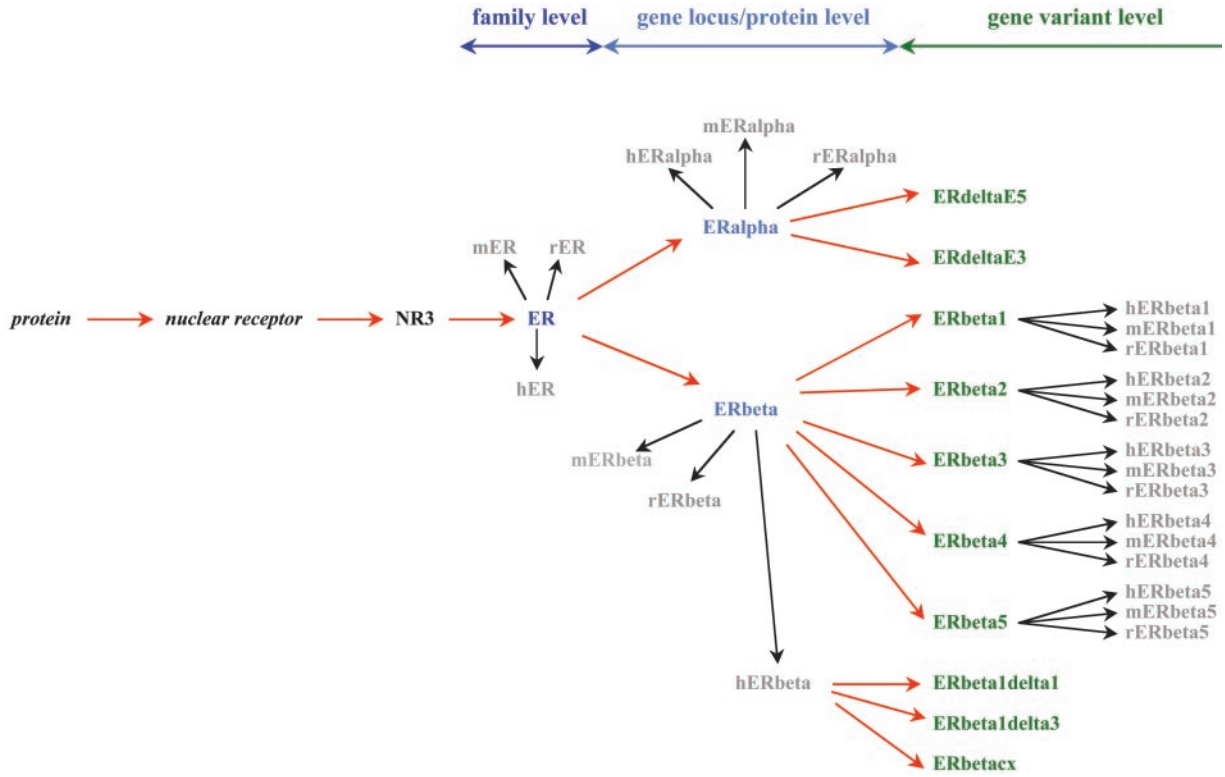

Fig. 6. Hierarchical Organization of the ER Family into Description Levels

The ER family is organized into different description levels: family level (*dark blue*), gene locus/protein level (*blue*), and gene variant level (*green*). The *arrows* represent "is-a" relationships between biological entities. NR3 is not a biological entity but a generic term issued from the NR nomenclature. The *red arrows* separate the different levels. The *black arrows* link each entity generally assigned to mammals to the corresponding entity in the following species: to human, mouse, or rat (*gray*). Assignment to species is available in the dictionary but was not used in the analysis of the data presented in this paper. hER, Human ER; mER, mouse ER; rER, rat ER.

1) First-letter case insensitivity for the term extraction (protein names, interaction terms, and species).

2) The selection of the longest term from the dictionary. For example, "ER" is not considered if the dictionary entry "ERβ" is also present in the sentence. In effect, this rules makes sure that the term with the highest resolution level is applied.

3) The avoidance of the "inclusion problem," which appears when one term from the dictionary is included in an unrelated longer term. For example, ER (estrogen receptor) and PR (progesterone receptor) are both included in the term "INTERPRETATION" that is found written with capital letters in many abstracts, but the extraction program should not consider "INTERPRETATION" as a string containing terms of interest. This rule allows, at the same time, retrieval in the text of new interesting protein names deriving from the dictionary's protein names (ERβ5 could be derived from ERβ, hER from ER) and recognition of the plural form of each protein name. The rules that permit the handling of the "inclusion problem" were set up to match terms only if they span whole words. This was done by requiring that a protein term must be found in a certain one-character context. The following contexts were chosen: 1) Protein names must not be preceded by a–z, A–Z, 0–9 or a dash (minus); 2) Depending on their last character, protein names must have the following context on the right as below:

Last character A–Z: anything but A–Z
Last character 0–9: anything but 0–9
Last character a–z: anything but a–z

In the last case, the terms without "s" as the last character were also matched if they were immediately followed by an "s" and only then by anything but a–z. No context was applied to verbs because they did not generate immediately obvious numbers of false positives.

Throughout the project, the extraction process was continuously refined and improved. The first important improvement concerning the trioccurrence extraction is the application of "stop lists" (see Fig. 1). These lists are used as an exclusion criterion for the terms they contain, as described below.

Stop List I. The trioccurrence extraction method has the side effect that search terms can be found in an unsuitable context: for example the term "binding" found in a string like "fatty-acid binding" or "binding to DNA" is of no interest for protein-protein interactions. The terms "X frequency" or "anti-X" or "binding of X to DNA," where X can be any of the dictionary's proteins, are also of no interest. To accelerate the curation process and improve its efficiency, a stop list containing nonuseful strings was built. At the end, this list was composed of approximately 1000 strings. If a word from the dictionary is included in one of these strings in the text, it should not be involved in any trioccurrence.

Stop List II. When considering protein names, it is necessary to handle acronyms (ER for "estrogen receptor" or VDR for "vitamin D receptor"). An acronym is generally short and then can have several significations: ER is an acronym for "estrogen receptor" but also for "endoplasmic reticulum." Thus, acronyms in connection with unwanted significations were entered into stop list II. This stop list also contains term pairs, including one term from the dictionary, which, according to our experience, were increasing the ratio of wrong interac-

tions (*e.g*. androgen receptor and carcinoma, estrogen receptor and metastasis, progesterone receptor and cancer). The program did not consider tri-occurrences when acronyms of interest are accompanied by unwanted significations or when terms from the dictionary are accompanied by unwanted terms at the sentence level.

Stop List III. This stop list contains protein pairs that were, according to our experience, never found to be involved in an interaction. For example, the pairs "estrogen receptor-androgen receptor" or "glucocorticoid receptor-progesterone receptor" are used to describe coexpression results, gene regulation, ligand binding to NRs, association with diseases, etc. but never a protein-protein interaction. Trioccurrences containing one of these couples and a term expressing an interaction were not considered and hence reduced the curation effort, at the expense of potentially missing some of these NR-NR interactions.

Table 2 shows the reduction in the number of trioccurrences and number of abstracts to be curated after applying each of the stop lists and a combination of the three lists. A limited sample of the extracted corpus was analyzed, out of which approximately one third of the trioccurrences and abstracts were eliminated from the curation process. This reduction is directly linked to the increase of the precision of the automated extraction.

The second major improvement concerning the trioccurrence extraction is the use of the dictionary's structure to select the appropriate interaction partner. For example, in some cases, abstracts contained inconsistencies of the following type: "Protein A binds to human estrogen receptor (ER)" or "Protein A binds to estrogen receptor (hER)." In fact, human estrogen receptor (or hER) is a child of estrogen receptor (or ER) in the dictionary. The interaction with protein A can only involve one of these two entities. The solution applied was to always favor the selection of the child at the expense of the parent for the trioccurrence extraction, and not to provide the curators with all the possible trioccurrences.

We defined the complete protein list of interest and queried MEDLINE for abstracts that contain at least one of the protein names. The resulting corpus of abstracts was then analyzed as follows. Abstracts were split into sentences, and in these sentences we marked, with an appropriate tag, words found in a list of terms (term list in Fig. 1) derived from the dictionary. The term list was obtained by applying a series of word-derivation rules as described above to three lists coming from the dictionary: protein 1 list, protein 2 list, and interaction terms. The final term list was obtained, after the application of the derivation rules, by the exclusion of all the strings contained in Stop List I. The text resulting from the tagging step contained only sentences with at least one trioccurrence where the dictionary terms were clearly marked. With LION's FSA (Finite State Automata) technology it was possible to filter 2.5 million selected abstracts for terms of the dictionary within a few hours. We eliminated the unwanted trioccurrences by applying Stop Lists II and III and finally extracted and dumped the ones to be curated into a relational database.

Curation Process

The Curation Interface. The curation interface allowed curators to select and curate the computer-extracted abstracts and trioccurences. In the curation process, a trioccurrence is selected, and the respective sentence is checked, as well as the abstract, when necessary. The curators check and correct the accuracy of the protein names in the dictionary. They can have access to the dictionary at any time. The curators can choose between different relation states for each trioccurrence. "Shows the relation" and "shows the negative relation" are chosen for a reported or denied interaction, respectively. The state "no relation shown" is used when there is no relation linking the three entities in the sentence. "Shows the relation, but not interesting" is used when an interaction involving proteins belonging to species other than mammal is stated. "No relation shown but true with respect to text mining" is chosen when the interaction between the entities is hypothetical (*e.g*. "our hypothesis was then A binds to B," "we investigated whether A binds to B," . . .) or when the

Table 2. Reduction in the Number of Trioccurrences (tri-occs) and Abstracts to be Curated after the Application of Stop Lists

	Purpose	Example	Total No. of tri-occs	Total No. of Abstracts	No. of tri-occs Excluded	No. of Abstracts Excluded
Without any Stop List	/	/	9970	4253	/	/
Stop List I only	To remove tri-occs where the term of interest is in a nonsuitable context	Binding to DNA	/	/	3080	1156
Stop List II only	To remove tri-occs when a nonsuitable term is found in the sentence	Endoplasmic reticulum (ER) instead of estrogen receptor (ER)	/	/	347	0
Stop List III only	To remove tri-occs containing protein couples never involved in an interaction	Estrogen receptor and androgen receptor	/	/	401	218
Stop Lists I, II, and III	To remove the maximum of false tri-occs	/	/	/	3828 = 38.4%	1374 = 32.3%

The stop lists were applied on a sample of 4253 abstracts from the extracted corpus.

interaction is not physical (functional interaction, synergistic interaction, interaction between signaling systems, interaction between genes, . . .). Note that the two latter stages are not of interest for building the nuclear receptors database, but to determine when the text-mining method applied has worked correctly, in a technical sense, when selecting these trioccurrences for the curation. The state "maybe" could be chosen by the curators, but it is changed to another state during the conflict resolution. "Duplication" is used to discard the trioccurrences added twice by mistake. The curators have the ability to create new trioccurrences in case the automatic program missed some (because proteins or interaction terms were not in the dictionary at the time of extraction) or, more frequently, when more than one sentence is needed for expressing an interaction. The string "implied interaction" can be used as an interaction term for manually added trioccurrences when a verb/noun is not precisely stated but the interaction between two proteins is obvious ("More recently, our lab has identified *ARA267*, a SET domain containing protein, and supervillin, an F-actin binding protein, as *AR* coregulators."). The *Results* chapter takes into consideration the fact that some of the available interactions could not be found by the automatic method but were added manually.

Resolution of Curation Conflicts. The results provided by two independent curators were merged, and the conflicting trioccurrences appeared as "unchecked" in the curation interface. A third biologist curated them, taking into consideration the relation states that the curators were choosing when the decision was uncertain.

Y2H Screening

The ligand binding domains of NRs were cloned into pGBT9 and transformed into yeast CG1945 using standard methods. All Y2H libraries used for screening were bought as pretransformed libraries in the yeast strain Y187 from CLONTECH (Palo Alto, CA). Culture and transformation of yeast cells were according to the instructions provided by CLONTECH. For screening, diploid cells containing both the NR and the library clones were generated by mating of yeast cells in Erlenmeyer flasks and selected for clones containing interacting hybrid proteins on selective medium lacking leucine, tryptophan, and histidine, containing 4-methyl umbelliferyl-α-D-galactoside (50 μM) and various amounts of 3-aminotriazole (Sigma, St. Louis, MO) in 96-well microtiter plates as described previously (32). Where appropriate, the ligand for the respective NR was added as indicated in the legend of Table 1. Positive cells were identified by measuring fluorescence at 460 nm (excitation at 365 nm) and passaged to new wells twice. Cells were then transferred to agar plates lacking leucine, tryptophan, and histidine using a manual 96-pin replicator and regrown before isolating the library insert via PCR using generic primers as recommended by CLONTECH. PCR products were resolved by agarose gel electrophoresis, and all reactions were collected where a single clear band was apparent. Inserts were sequenced at GATC Biotech AG (Konstanz, Germany) and analyzed by sequence comparison to public databases using the BLAST algorithm (33). Clones that corresponded to untranslated regions or the noncoding strand were discarded. All compounds were from Sigma.

Nomenclature

All NR names refer to the official NR nomenclature (34).

Web Site References

Web site references are as follows: http://www.ncbi.nlm.nih.gov/Entrez, Entrez search and retrieval system homepage; http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html, Entrez progamming utilities; http://srs.ebi. ac.uk, SRS entry page at the European Bioinformatics Institute.

Acknowledgments

We thank Ralf Tolle, Ingo Kober, and Jan Mous for critical reading of the manuscript; Harald Kranz and Markus Post for cloning and sequencing support; Eva Löser, Simone Scheurer, Rainer Kern, and Jörg Suckow for their work with the Y2H screens; and all colleagues at LION for providing an enthusiastic working environment.

Received December 17, 2002. Accepted April 28, 2003.

Address all correspondence and requests for reprints to: Manfred Koegl, Phenex Pharmaceuticals AG, Im Neuenheimer Feld 515, 69120 Heidelberg, Germany. E-mail: manfred.koegl@phenex-pharma.com.

This work was supported by the Department of Biotechnology from the German Ministry of Research (BmBF-Project 0312385).

Current address for M.A.: Phenex Pharmaceuticals AG, Im Neuenheimer Feld 515, 69120 Heidelberg, Germany.

Current address for B.H.: Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Current address for D.R.-S. and H.K.: EMBL-Outstation Hinxton, Hinxton, Cambridge CB10 1SD, United Kingdom.

The retrieved data are available online (published as supplemental data on The Endocrine Society's Journals Online web site at http://mend.endojournals.org.

REFERENCES

1. Aranda A, Pascual A 2001 Nuclear hormone receptors and gene expression. Physiol Rev 81:1269–304
2. McKenna NJ, O'Malley BW 2002 Minireview: nuclear receptor coactivators—an update. Endocrinology 143: 2461–2465
3. Dilworth FJ, Chambon P 2001 Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription. Oncogene 20:3047–3054
4. Kliewer SA, Lehmann JM, Willson TM 1999 Orphan nuclear receptors: shifting endocrinology into reverse. Science 284:757–760
5. Chan S 2002 A review of selective estrogen receptor modulators in the treatment of breast and endometrial cancer. Semin Oncol 29:129–133
6. Bentrem DJ, Craig Jordan V 2002 Tamoxifen, raloxifene and the prevention of breast cancer. Minerva Endocrinol 27:127–139
7. Boyack M, Lookinland S, Chasson S 2002 Efficacy of raloxifene for treatment of menopause: a systematic review. J Am Acad Nurse Pract 14:150–165
8. Takeyama K, Masuhiro Y, Fuse H, Endoh H, Murayama A, Kitanaka S, Suzawa M, Yanagisawa J, Kato S 1999 Selective interaction of vitamin D receptor with transcriptional coactivators by a vitamin D analog. Mol Cell Biol 19:1049–1055
9. Kodera Y, Takeyama K, Murayama A, Suzawa M, Masuhiro Y, Kato S 2000 Ligand type-specific interactions of peroxisome proliferator-activated receptor γ with transcriptional coactivators. J Biol Chem 275:33201–33204
10. Kraichely DM, Sun J, Katzenellenbogen JA, Katzenellenbogen BS 2000 Conformational changes and coactivator recruitment by novel ligands for estrogen receptor-α and estrogen receptor-β: correlations with biological charac-

ter and distinct differences among SRC coactivator family members. Endocrinology 141:3534–3545
11. Bramlett KS, Wu Y, Burris TP 2001 Ligands specify coactivator nuclear receptor (NR) box affinity for estrogen receptor subtypes. Mol Endocrinol 15:909–922
12. Bramlett KS, Burris TP 2002 Effects of selective estrogen receptor modulators (SERMs) on coactivator nuclear receptor (NR) box binding to estrogen receptors. Mol Genet Metab 76:225–233
13. McDonnell DP, Connor CE, Wijayaratne A, Chang CY, Norris JD 2002 Definition of the molecular and cellular mechanisms underlying the tissue-selective agonist/antagonist activities of selective estrogen receptor modulators. Recent Prog Horm Res 57:295–316
14. Duarte J, Perriere G, Laudet V, Robinson-Rechavi M 2002 NUREBASE: database of nuclear hormone receptors. Nucleic Acids Res 30:364–368
15. Horn F, Vriend G, Cohen FE 2001 Collecting and harvesting biological data: the GPCRDB and NucleaRDB databases. Nucleic Acids Res 29:346–349
16. Martinez E, Moore DD, Keller E, Pearce D, Vanden Heuvel JP, Robinson V, Gottlieb B, MacDonald P, Simons Jr S, Sanchez E, Danielsen M 1998 The Nuclear Receptor Resource: a growing family. Nucleic Acids Res 26: 239–241
17. Mack R, Hehenberger M 2002 Text-based knowledge discovery: search and mining of life-sciences documents. Drug Discov Today 7:S89–S98
18. Blaschke C, Hirschman L, Valencia A 2002 Information extraction in molecular biology. Brief Bioinform 3:154–165
19. Hahn U, Romacker M, Schulz S 2002 Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 7:338–349
20. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L 2000 EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 5:516–525
21. Raychaudhuri S, Schutze H, Altman RB 2002 Using text analysis to identify functionally coherent gene groups. Genome Res 12:1582–1590
22. Blaschke C, Andrade MA, Ouzounis C, Valencia A 1999 Automatic extraction of biological information from scientific text: protein-protein interactions. Proc International Conference on Intelligent Systems for Molecular Biology (ISMB), 1999, Heidelberg, Germany, pp 60–67
23. Marcotte EM, Xenarios I, Eisenberg D 2001 Mining literature for protein-protein interactions. Bioinformatics 17: 359–363
24. Oltvai ZN, Barabasi AL 2002 Systems biology life's complexity pyramid. Science 298:763–764
25. Ideker T, Galitski T, Hood L 2001 A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372
26. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L 2001 Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934
27. Fields S, Song O 1989 A novel genetic system to detect protein-protein interactions. Nature 340:245–246
28. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
29. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415: 180–183
30. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW 2001 BIND—The Biomolecular Interaction Network Database. Nucleic Acids Res 29:242–245
31. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, 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
32. Motz M, Kober I, Girardot C, Loeser E, Bauer U, Albers M, Moeckel G, Minch E, Voss H, Kilger C, Koegl M 2002 Elucidation of an archaeal replication protein network to generate enhanced PCR enzymes. J Biol Chem 277: 16179–16188
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990 Basic local alignment search tool. J Mol Biol 215: 403–410
34. 1999 A unified nomenclature system for the nuclear receptor superfamily. Cell 16:161–163