

Searching for Meaningful Feature Interactions with Backward-Chaining Rule Induction

Doug Fisher¹, Mary Edgerton^{2,3}, Lianhong Tang⁴,
Lewis Frey³, and Zhihua Chen¹

¹ Department of Electrical Engineering and Computer Science, Vanderbilt University,
Nashville, TN, USA 37235

{douglas.h.fisher, mary.edgerton, l.tang, lewis.j.frey,
zhihua.chen} @vanderbilt.edu
<http://www.vuse.vanderbilt.edu/~dfisher/>

² Department of Pathology, Vanderbilt University Medical Center

³ Department of Biomedical Informatics, Vanderbilt University Medical Center

⁴ Vanderbilt Ingram Cancer Center, Vanderbilt University Medical Center

Abstract. Exploring the vast number of possible feature interactions in domains such as gene expression microarray data is an onerous task. We propose Backward-Chaining Rule Induction (BCRI) as a semi-supervised mechanism for biasing the search for plausible feature interactions. BCRI adds to a relatively limited tool-chest of *hypothesis generation* software, and it can be viewed as an alternative to purely unsupervised association rule learning. We illustrate BCRI by using it to search for gene-to-gene causal mechanisms. Mapping hypothesized gene interactions against a domain theory of prior knowledge offers support and explanations for hypothesized interactions, and suggests gaps in the current domain theory, which induction might help fill.

1 Introduction

With the increasing investment in gene expression microarray technology, there has been a move toward a “systems biology” approach to understanding the coupling of gene networks and signaling cascades that describe the phenotypes of living matter (e.g., [1],[2],[3]). This has led to a call for tools to (semi-)automatically explore the space of genomic interactions (e.g., [4]) in order to reduce the set of interactions to a manageable set for examination. The goal of this exploration is to focus analysts on plausible interactions, pathways, and markers, which can then be scrutinized further with hypothesis testing methods.

Consistent with research on other exploratory strategies (e.g., [5],[6],[7],[8]), we describe an investigation of *backward-chaining rule induction* (BCRI) for hypothesizing molecular causality and functional interactions from gene expression microarray data. BCRI is a novel strategy for restricting the search through a rule-space to those rules with traceable influence on a given top-level target class. Put simply, BCRI is given a top-level classification with labeled data, and rule induction is performed to find rules that predict the specified class. Antecedent conditions found in discovered rules then become “sub-goals”, and rule induction is repeated on the data using these

sub-goal conditions as classes. The process of backward-chaining on rule antecedent conditions is repeated until a termination condition is satisfied.

BCRI is intermediate between supervised rule induction and unsupervised rule induction (e.g., association rule learning). Rather than an unconstrained exploration of the space of associations between variables, as would occur in association rule learning [9], only associations that are weakly tied to a top-level class are examined. While BCRI’s search through association space will miss many associations (with any given top-level class), we expect that the density of “interesting” rules that it discovers will be higher than if uncovered by standard association rule learning, though this paper does not test this hypothesis directly.

BCRI can be viewed as one component in a process of *iterative exploration*. Induction from data (e.g., BCRI) can be used to find plausible interactions, which are then compared against prior knowledge. Prior knowledge can be used to (1) explain plausible interactions found through induction, (2) filter or rank these possibilities for an analyst (e.g., interactions that are already well-established in the literature might be ranked low, as might be those in which prior knowledge offers too few constraints on possible explanations), (3) implicate additional features or suggest pruning “redundant” features for subsequent induction (e.g., feature selection), (4) reveal gaps in current knowledge that induction may help fill. We look at examples of this last case in **Section 3**.

Our contributions are (1) the definition of the BCRI task abstraction, (2) the implementation of an initial prototype of BCRI, which we call C45-BCRI, (3) an illustration of BCRI in the domain of cancer prognosis, and (4) a demonstration of how BCRI generated hypotheses (e.g., gene interactions) may help fill gaps in prior knowledge. In **Section 2** we describe our implementation of BCRI and report our results in the domain lung cancer prognosis from clinical and gene expression data. In **Section 3** we match selected rules against prior knowledge in the form of an established gene interaction network. Inductively derived rules suggest values for gaps in current knowledge and suggest other plausible hypotheses. **Section 4** closes with a discussion on automating aspects of iterative exploration by coordinating the application and derivation of domain and induced knowledge.

2 Backward-Chaining Rule Induction (BCRI)

To summarize, the initial step of BCRI builds decision rules for predicting a user-specified class or outcome. The antecedents of rules discovered in this first step then become outcomes for which decision rule models are constructed in the second step. Antecedents of rules found in this second step, then become outcomes for decision rule learning in the third step, and so on.

As an illustrative example, in this paper we apply BCRI to published gene-expression and clinical data from lung cancer patients [10]. The data contains 61 instances defined over 4,996 gene attributes and eleven clinical attributes (5007 total). Classification as *High versus Low* risk is the as a top-level task that “kick-starts” BCRI in our application. For our analysis, patients who died at 30.1 months or less following diagnosis are high risk, and others are low risk.

We distinguish the general BCRI task abstraction from our initial implementation of BCRI. We implement BCRI as a wrapper around a rule-induction engine, which is illustrated in **Table 1** with pseudo-C code (local variable declarations excluded). BCRI is passed the labeled data, a set of the target classes used to label the data, and three functions: *RuleInducer*, *PriorityFn*, and *TerminateFn*.

Table 1. Pseudo-C for Backward-Chaining Rule Induction

```

RuleSet BCRI (DataSet Data,
              TargetSet Classes,
              RuleSet (* RuleInducer) (DataSet, TargetCondition),
              float (* PriorityFn)(Rule),
              int (* TerminateFn) (Rule)) {
  PQ = InitializePriorityQueue(PriorityFn);
  FOR each class in Classes, Enqueue(PQ, [class → ___]);
  WHILE (NOT Empty(PQ)) {
    R = Dequeue(PQ); /* and place R in Results SET*/
    IF (NOT (* TerminateFn)(R) {
      FOR each a IN ANTECEDENTS(R) {
        Children = (* RuleInducer) (Data, a);
        FOR each c IN Children Enqueue(PQ, c)
      }
    }
  } /* end WHILE */
} /* end BCRI */

```

RuleInducer can, in principle, be any supervised rule discovery system that, given a class, will return rules that predict that class (i.e., *RuleInducer* is not a classifier per se as no rules predicting the complement of class are explicitly returned). Parameters shown for *RuleInducer* might be changed in minor ways to support differing induction engines. Our current implementation uses C4.5-rules [11], which first builds a decision tree to discriminate the values of a dependent attribute (i.e., C4.5-rules builds the classifier), then converts the tree to a set of rules. We do not detail the process here, as it is well established in the literature. We use the standard defaults for C4.5-rules. Moreover, as stated, the wrapper model that we have implemented assumes that *RuleInducer* discovers rules whose consequents are all of the same class. Thus, while C4.5-rules would discover rules for the complement of a class as well, we filter these out. The BCRI prototype is not optimal in terms of cost, but it allows us to investigate the BCRI methodology by exploiting a well-established rule learning algorithm. In the remainder of this section, we will refer to this C4.5-rules procedure, with filtering of complement rules, as simply C4.5.

PriorityFn is applied to a rule and returns a score. This score is used to store the rule on a priority queue of other scored rules. In our current implementation, the coverage of the rule (i.e., the number of instances in the data that satisfy the rule’s antecedent), is used to organize the priority queue. Other possibilities include the rule’s accuracy or the like. Our choice of coverage, versus accuracy or a like measure, is motivated by the observation that rule-learning systems tend to produce accurate rules (relative to a data specific upper bound), but that these rules vary significantly in coverage. We prefer to favor rules that cover a large proportion of data.

TerminateFn returns 0/1, indicating whether a rule should be further expanded (i.e., continued backward chaining on its antecedents). Currently, we implement a depth bound and only backward chain a specified number of levels. Other strategies include specifying a minimal coverage or confidence bound.

C45-BCRI, which includes a wrapper around C4.5 as just described, is what we call this paper’s implementation of BCRI. Using High and Low Risk as the top-level classification, C45-BCRI begins with (Risk=Low) (cov 42/61) and (Risk=High) (cov 19/61) placed on the priority queue (i.e., passed as Classes to BCRI). The term “cov” is an abbreviation for data coverage (described above in *PriorityFn*) for the condition just described.

(Risk=Low) is dequeued. Application of C4.5-rules yields a single rule, which is placed on the queue:

[(Stage=1) \rightarrow (Risk=Low) (cov 48/61) \parallel (Risk=High) \rightarrow (cov 19/61)].

(Stage=1) \rightarrow (Risk=Low) is dequeued and C4.5-rules yields a rule, which is added to the queue:

[(ELA2 > 163.3) \rightarrow (Stage 1) (cov 45/61) \parallel (Risk = High \rightarrow (cov 19/61)]

The first of these rules is dequeued. A new rule is learned:

(MRPL19<= 161.4) & (EIF2S1 > 52) & (KRT15 <= 616.8) \rightarrow (ELA2 >163.3)
(cov: 45/61)

This rule is queued, resulting in the following priority queue:

[(MRPL19<= 161.4) & ... \rightarrow (ELA2 >163.3) (cov: 45/61) \parallel (Risk = High \rightarrow (cov 19/61)]

Having the highest priority (coverage), this same rule is immediately dequeued. Each individual antecedent serves in turn as a class for rule induction. A simple depth bound is used to terminate backward chaining, and these labeled rules are terminal.

Table 2 shows the 19 rules learned from the lung cancer data by backward chaining with C45-BCRI to a depth of three, beginning with an initial queue of

[(Risk = Low) \rightarrow \parallel (Risk = High) \rightarrow]

The ordering of rules is not strictly indicative of the order in which they were discovered. Rule number is given with its associated depth in the backward chaining

process. Indentation indicates a parent child relationship. “acc” denotes accuracy (percent correct predictions) and “cov” denotes coverage, which is the number of cases satisfying the antecedents over the total number of samples.

The network of rules learned by BCRI (C45-BCRI and otherwise) is an AND/OR graph, much like the rule bases of expert systems such as Mycin [12]. We do *not* discuss the inference possibilities of such networks from an expert-system perspective. Rather, our primary goal here is a limited, focused exploration of the associations between variables, which is directly (initially) or indirectly (as backward chaining proceeds) tied to top-level class(es).

3 Hypothesized Pathways from BCRI and Prior Knowledge

BCRI rules can be used to find plausible interactions, which can then be compared against prior knowledge. Prior knowledge can be used to (1) explain plausible interactions found through induction, (2) filter or rank these possibilities for an analyst (e.g., interactions that are already well-established in the literature might be ranked low, as might be those in which prior knowledge offers too few constraints on possible explanations), (3) implicate additional features or suggest pruning “redundant” features for subsequent induction (e.g., feature selection), (4) reveal gaps in current knowledge that induction may help fill. We look at examples of this last case.

We will focus here on C45-BCRI rules that are reflected in existing knowledge, or pathways where we can make inferences from a combination of induced and existing knowledge. For each C45-BCRI rule we used Pathway Assist™ [13] to build the shortest pathway known from prior knowledge (as encoded in Pathway Assist™) between gene expression attributes of the rule. We have also used PubMed [14], LocusLink ([15],[16]), and GeneCards [17] as sources for peer-reviewed literature, chromosomal location, and functional annotation, respectively.

The type of interaction found in Pathway Assist™ pathways may involve gene expressions, which are the measured quantities in the data we used, or it may involve the protein product of gene expression, which is not measured but *might* be inferred from the gene expression level. We say “might” because gene expression quantities are not always directly proportional to protein product concentrations secondary to other factors affecting protein concentration (e.g. degradation, export, etc.)

Example 1: Our first example is one where a C45-BCRI-discovered rule is reflected by existing knowledge. In Rule 8./1, we have

8./1 (ELA2 \leq 163.3) & (SERPINA1 > 65) \rightarrow (Stage=3) [acc: 89.1% cov: 12/61]

Existing knowledge from Pathway Assist™, illustrated in Figure 1, gives us relationships between ELA2 and SERPINA1, where the protein products are indicated by the large ovals, a binding interaction is indicated by the dot relationship between the ovals, and regulation is indicated by a square along a dotted line. Table 3 describes the type of reaction between specific nodes, the nodes themselves, and the effect of one node upon the other using the direction indicated in the nodes list.

Table 2. Rules induced by C45-BCRI

(Risk=Low) →

Rule # /Depth

- 1./0 (Stage=1) → (Risk=Low)
- 2./1 (ELA2 > 163.3) → (Stage=1) [acc: 94.4% cov: 45/61]
- 3./2 (MRPL19 ≤ 161.4) & (EIF2S1 > 52) & (KRT15 ≤ 616.8) → (ELA2 > 163.3) [acc: 97.0% cov: 45/61]
- 4./3 (TRIP12 ≤ 1176) & (NAPG ≤ 243) → (MRPL19 ≤ 161.4) [acc: 97.4% cov: 53/61]
- 5./3 (FXN > 37.8) → (EIF2S1 > 52) [acc: 97.6% cov: 57/61]
- 6./3 (CTRL > 194.4) & (IDS ≤ 163.3) → (KRT15 ≤ 616.8) [acc: 97.5% cov: 54/61]

(Risk=High) →

Rule # /Depth

- 7./0 (Stage=3) → (Risk=High)
- 8./1 (ELA2 ≤ 163.3) & (SERPINA1 > 65) → (Stage=3) [acc: 89.1% cov: 12/61]
- 9./2 (MRPL19 > 161.4) → (ELA2 ≤ 163.3) [acc: 84.1% cov: 8/61]
- 10./3 (TRIP12 > 1176) → (MRPL19 > 161.4) [acc: 75.8% cov: 5/61]
- 11./3 (NAPG > 243) → (MRPL19 > 161.4) [acc: 70.7% cov: 3/61]
- 12./2 (KRT15 > 616.8) → (ELA2 ≤ 163.3) [acc: 79.4% cov: 5/61]
- 13./3 (CTRL ≤ 194.4) → (KRT15 > 616.8) [acc: 70.7% cov: 4/61]
- 14./3 (IDS > 163.3) → (KRT15 > 616.8) [acc: 45.3% cov: 3/61]
- 15./2 (PLAB ≤ 3703.9) & (H3FD ≤ 167.6) & (ANXA5 > 750) & (DDX5 ≤ 2804.7) → (SERPINA1 > 65) [acc: 96.9% cov: 44/61]
- 16./3 (KIAA0618 > 27.2) → (PLAB ≤ 3703.9) [acc: 95.5% Data cov: 54/61]
- 17./3 (SC4MOL > 32) → (H3FD ≤ 167.6) [acc: 97.5% cov: 54/61]
- 18./3 (AKAP > 496) & (SLC14A2 ≤ 397.1) → (ANXA5 > 750) [acc: 97.4% cov: 53/61]
- 19./3 (KRT13 ≤ 262.9) → (DDX5 ≤ 2804.7) [acc: 97.6% cov: 57/61]

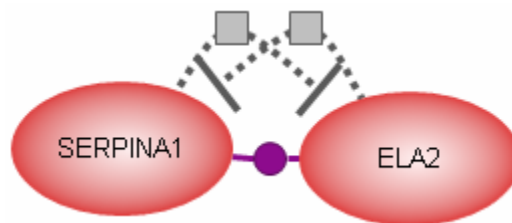


Fig. 1. Pathway Assist™ diagram showing SERPINA1 and ELA2 relationships of Example 1

Table 3. Details of Example 1 relationships given by Pathway Assist™

| Type | Nodes | Effect |
|------------|--------------------|----------|
| Binding | ELA2 ---- SERPINA1 | |
| Regulation | ELA2 --- SERPINA1 | negative |
| Regulation | SERPINA1 --- ELA2 | negative |

There are three known interactions between ELA2 and SERPINA1. Their protein products bind together, the protein product of ELA2 gene expression inhibits (i.e., Effect is negative) the gene expression of SERPINA1, and the protein product of SERPINA1 inhibits the gene expression of ELA2 (reciprocal down regulation).

The C45-BCRI rule suggests that ELA2 and SERPINA1 are also coupled by reciprocal down regulation of gene expression. The reciprocal negative regulation effect (down regulation) is reflected in the opposing relative quantities of the attributes in the antecedent of the rule, i.e., that ELA2 is depressed below a value and SERPINA1 is elevated above a value.

The regulatory relationship indicates that SERPINA1 will be highly expressed, ELA2 expression will be depressed. We can hypothesize that the activity of the protein product of ELA2, which is inhibited by binding with the protein product of SERPINA1, will be low as a condition for a high risk tumor. In a 1992 study of adenocarcinomas, high levels of alpha-1-antitrypsin, the protein product of SERPINA1 were found to be associated with higher stage disease [18]. However, high levels of elastase, the protein product of ELA2, in lung tumor tissue has also been correlated with higher stage tumors and poor survival in patients with lung cancer ([19],[20]).

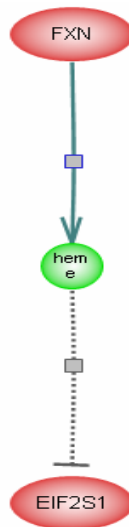


Fig. 2. Pathway Assist™ diagram illustrating FXN and EIF2S1 relationships of Example 2

Table 4. Details of Example 2 relationships given by Pathway Assist™

| Type | Nodes | Effect |
|--------------|------------------|----------|
| Regulation | heme --- EIF2S1 | negative |
| MolSynthesis | FXN ---> heme | unknown |

We learn from BCRI in combination with existing knowledge that we may need to study the interaction of elastase and alpha-1-antitrypsin, and not either of these in isolation, to understand their role in lung cancer survival.

Example 2: As a second example, consider Rule 5./3,

$$5./3 \quad (FXN > 37.8) \rightarrow (EIF2S1 > 52) \quad [\text{acc: } 97.6\% \text{ cov: } 57/61]$$

Pathway Assist™ shows that FXN has an “unknown” effect on the molecular synthesis of heme, the interaction represented as a solid line with a square in Figure 2, and that heme, a small molecule depicted by the small, central oval, inhibits the gene expression of EIFS2. Relational details are listed in Table 4.

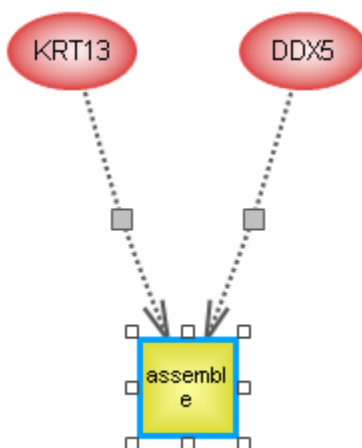


Fig. 3. Pathway Assist™ diagram of KRT13 and DDX5 relationships of Example 3

From our C45-BCRI rule, if we accept that elevated gene expression of FXN leads to elevated levels of its protein product frataxin, then we can *infer* that frataxin *blocks* the molecular synthesis of heme to results in elevated expression of EIFS2. Thus, the inductively derived rule, which might be tentatively abstracted as (FXN → EIF2S1, effect positive), together with (heme--|EIF2S1, effect negative) from prior knowledge, suggests that (FXN→ heme, effect *negative*, in place of unknown). This example illustrates where *induction can suggest fillers for gaps in background knowledge*.

Example 3: In Rule 19./3, 23 have from C45-BCRI

$$19./3 \quad (KRT13 \leq 262.9) \rightarrow (DDX5 \leq 2804.7) \quad [\text{acc: } 97.6\% \text{ cov: } 57/61]$$

Table 5. Details of relationships of Example 3 given by Pathway Assist™

| <u>Type</u> | <u>Nodes</u> | <u>Effect</u> |
|-------------|---------------------|---------------|
| Regulation | KRT13 ---> assemble | unknown |
| Regulation | DDX5 ---> assemble | unknown |

and from Pathway Assist™ we have the diagram of Figure 3 and description of interactions in Table 5. The square labeled “assemble” represents a cell function of assembly, for example assembling a scaffold of filamentous proteins either into a structure for cell shape, or a scaffold upon which catalyzed reactions can take place.

From GeneCards and Locus Link, we learn that the protein product for KRT13 is keratin 13, a cytoskeletal protein that functions in maintaining the integrity of the cell shape and may also function as a support structure in cell reactions. p68 RNA helicase is the protein product for DDX5 and functions as an RNA-dependent ATPase (provides energy by breaking down ATP). Its presence in the nucleus is an indicator of proliferation, which is an important process in cancer.

The gene for KRT13 is located at chromosome position 17q12 to 17q21.2, while DDX5 is located at 17q21. This suggests that the transcription of DDX5 is associated with transcription of KRT13. Interestingly, Massion and Carbone [21] describe amplifications (increased numbers of copies of genes) in the 17q region of the genome (chromosome 17) that are associated with lung cancer.

From Pathway Assist™, we see that both the protein product of KRT13 and of DDX5 have an unknown role in assembly. Given the proximity of their locations on chromosome 17q, their common, though with unknown effect, role in assembly, and the correlation of amplification of 17q with lung cancer, we infer that 1) the genes on 17q that have a role in lung cancer include KRT13 and DDX5, 2) KRT13 and DDX5 are regulated by a common factor which controls transcription of the two together, 3) *the effect of KRT13 and DDX5 on assembly is the same* (either both positive or both negative), and 4) that the assembly process promotes proliferation.

4 Concluding Remarks

BCRI has been suggested as a means of biasing the search for gene interactions, and feature interactions generally. In particular, our contributions are (1) the definition of the BCRI task abstraction, (2) the implementation of an initial prototype of BCRI, which we call C45-BCRI, (3) an illustration of BCRI in the domain of cancer prognosis, and (4) an illustration of how (C45-)BCRI generated rules (e.g., gene interactions), coupled with prior knowledge, suggest hypotheses about the ways that genes interact that is not yet established in the literature.

Example 2 (Figure 2, Table 4), in particular, is a good example of a domain-independent hypothesis generation strategy. In this example, constraints from background knowledge and a C45-BCRI rule were sufficient to suggest, through qualitative reasoning, the value of an unknown effect. Of course, this reasoning only suggests hypotheses, but at least one other example of this possibility is found in our data. One direction of future research into iterative exploration is (semi-)automate the process of (1) examining a BCRI rule, (2) bolstering confidence in a suggested ef-

fect/correlation (positive, negative) through additional inductive means, (3) matching this rule against prior knowledge, (4) qualitatively reasoning about what can be inferred from the prior and inductive knowledge. This direction of research is related to work in scientific discovery (e.g., [22]) and theory revision (e.g., [23]).

A second direction is to use what is learned by what is found in prior knowledge to bias subsequent induction through feature selection with theory-derived features (e.g., [24]). We are pursuing theory driven feature selection strategies elsewhere ([25],[26]).

Finally, our only implementation of BCRI uses C4.5 as the core rule induction engine. C4.5 is biased to greedily find a minimal number of rules. A better choice as the base rule-induction engine may be a method such as Brute [27], which more extensively searches the space of rules, and generally returns many more rules. This latter characteristic will further motivate and necessitate work into using background knowledge to filter/rank hypothesized interactions for expert commentary.

Acknowledgements

We thank the reviewers for helpful and corrective comments. We also thank Mark Ross for his help in building and maintaining hardware on which BCRI was implemented. The research efforts of D.F., M.E., L.T., and Z.C. are supported in part by a National Institute of Health (NLM) grant (1R01LM008000) to M.E. and by funds from the Office of Research at Vanderbilt University Medical Center. L.F. is supported by a fellowship from the National Library of Medicine (5T15LM007450).

References

1. Guffanti, A. (2002). Modeling molecular networks: a systems biology approach to gene function. *Genome Biol* 3: reports4031.
2. Weston, A., & Hood, L. (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 3:179-96.
3. Provar, N., & McCourt, P. (2004). Systems approaches to understanding cell signaling and gene regulation. *Curr Opin Plant Biol* 7:605-9.
4. Huels, C., Muellner, S., Meyer, H., et al. (2002). The impact of protein biochips and microarrays on the drug development process *Drug Discov Today* 7(18 Suppl):S119-24.
5. Evans, B., & Fisher, D. (1994). Overcoming process delays with decision tree induction. *IEEE Expert* 9: 60-66.
6. Evans B., & Fisher, D. (2002). Decision tree induction to minimize process delays. In *Handbook of Data Mining and Knowledge Discovery*, W. Klossgen & J. Zytkow (Eds). Oxford, UK. Oxford University Press. pp 874-881.
7. Waitman, L.R., Fisher, D., & King, P. (2003). Bootstrapping rule induction. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society Publications Office, Los Alamitos CA, pp. 677-680.
8. Waitman, L.R., Fisher, D., & King, P. (in press) Bootstrapping rule induction to achieve and increase rule stability. *Journal of Intelligent Information Systems*.
9. Mannila, H. (2002). Association rules. In *Handbook of Data Mining and Knowledge Discovery*, W. Klossgen & J. Zytkow (Eds). Oxford, UK. Oxford University Press. pp 344-348.

10. Beer, D, Kardia, S, Huang, C, et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 8: 816-824.
11. Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Francisco. Morgan Kaufmann. URL: <http://quinlan.com>
12. Shortliffe, E., Davis, R., Axline, S., et al (1975). Computer –based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comput. Biomed. Res* 8: 303-320.
13. Nikitin, A., Egorov, S., Daraselia, N., & Mazo, I. (2003). Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics*, 19: 2155-2157.
14. PubMed Central, a free archive of life sciences journals. URL: <http://www.pubmedcentral.nih.gov/>
15. Pruitt, K., Katz, K., Sicotte, H. et al. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 16(1):44-47.
16. URL: <http://www.ncbi.nlm.nih.gov/projects/LocusLink/>
17. Pruitt, K., & Maglott, D. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29(1):137-140.
18. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., et al. (1997). GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel) URL: <http://bioinformatics.weizmann.ac.il/cards>
19. Higashiyama, M., Doi, O., Kodama, K., et al. (1992). An evaluation of the prognostic significance of alpha-1-antitrypsin expression in adenocarcinomas of the lung: an immunohistochemical analysis. *Br J Cancer* 65: 300-302.
20. Yamashita, J., Tashiro, K., Yoneda, S., et al. (1996). Local increase in polymorphonuclear leukocyte elastase is associated with tumor invasiveness in non-small cell lung cancer. *Chest* 109: 1328-1334.
21. Yamashita, J., Ogawa, M., Abe, M., et al (1997) Tumor neutrophil elastase is closely associated with the direct extension of non-small cell lung cancer into the aorta. *Chest* 111:885-90.
22. Massion, P., & Carbone, D. (2003). The molecular basis of lung cancer: molecular abnormalities and therapeutic implications. *Respiratory Research* 4: 12.
23. Langley, P., Shrager, J., & Saito, K. (2002). Computational discovery of communicable scientific knowledge. In “Logical and Computational Aspects of Model-Based Reasoning” L. Magnani, N.J. Nersessian, & C. Pizzi (Eds). Kluwer.
24. Mooney, R. (1993). Induction over the unexplained: Using overly-general theories to aid concept learning. *Machine Learning* 10: 79-110.
25. Ortega, J., & Fisher, D. (1995). Flexibly exploiting prior knowledge in empirical learning. In Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Francisco, pp. 1041-1047.
26. Frey, L., Edgerton, M., Fisher, D., Tang, L., & Chen, Z. (2005). Discovery of molecular markers of poor prognosis from rule induction methods. Poster presented at the American Association for Cancer Research (AACR) Conference on Molecular Pathogenesis of Lung Cancer: Opportunities for Translation to the Clinic (San Diego, CA).
27. Frey, L., Edgerton, M., Fisher, D., Tang, L., & Chen, Z. (under review). Using prior knowledge and rule induction methods to discover molecular markers of prognosis in lung cancer. American Medical Informatics Association Symposium 2005 (Washington DC).
28. Riddle, P., Segal, R., & Etzioni, O. (1994). Representation Design and Brute-force induction in the Boeing Manufacturing Domain. *Applied Artificial Intelligence* 8: 125-147.