Abstract

This is a manual describing how human coders should annotate a corpus of genetic counseling patient letters such as the sample letter in (Baker et al., 2002). The purpose of annotating the letters is to provide information useful for developing broad coverage computerized tools for genetics counselors.

Types of Annotation Schemes

The letters will be annotated using two different annotation schemes, each with different goals:

1. **Genetic counseling topic outline**: this is an indexing of topic headings similar to those in the outline given in (Baker et al., 1998, p. 238). Note that our main reason for indexing the topics is to reveal the contexts in which different writing patterns are used.

2. **Writing patterns**: The purpose of this part of the analysis is to identify the major patterns of information presentation in the text. These patterns can be identified independently of the specific genetics information that they are used to present. The patterns of primary interest are those used to (1) present or justify hypotheses (e.g. possible diagnoses), actions (e.g. diagnostic tests), and predictions (e.g. recurrence risks) on the basis of observations (e.g. family history or test results), (2) address known or potential misunderstandings of the client, and (3) help the client overcome emotional barriers impeding comprehension.

Other annotation schemes developed for analysis of genetic counseling sessions (Kessler, 1979; Liede, 2000; Resta, 2000) identify interactional features of the counselor-client dialogue. Currently, we do not plan to apply this type of scheme to the letters.

General Procedure for Coding Letter

This assumes that you have successfully completed the training process. Before you begin coding a letter, each sentence in it will be numbered sequentially and the letter will be presented in a tabular form (called the **coding table**). In addition, any information that could reveal the identity of actual clients has been removed or replaced with a generic term typed in upper case and given in square brackets, for example, [PROBAND]. The description of the procedure you should follow has been broken down into the following steps. However, you are not required to do the steps in this order and you may return to any step as often as necessary.
1. Read the letter to **understand its full meaning**. You may read it more than once and consult genetics reference materials if necessary to help interpret it. Your analysis of the letter in steps 2-5 should be based upon what you think the **writer meant, presumed, or implied**, i.e., do not limit your interpretation to the writer's exact words!

2. Assign the **topic outline headings**. See Appendix A for detailed instructions.

3. Annotate **writing pattern tags** in the text. These tags identify words or phrases that may play a role in the writing patterns that you will identify in the following steps. See Appendix B for detailed instructions.

4. Draw one or two **network graphs** showing relationships between certain writing pattern tags that you found in the text. See Appendix C for instructions.

5. Describe the **probability statements** in the text based on the writing pattern tags that you found in the text. See Appendix D for instructions.

6. Keep a record of any problems or questions that you have about the procedure that were not resolved by the time that you finished coding the letter.

A fully annotated letter (including coding table and network graphs) is given as an example in Appendix E.

**References**


**Appendix A. Topic Outline Headings**

Assign one or more paragraph section numbers to each sentence, using the following table of topic headings. Most likely, the paragraphs in the letter to be coded will not exactly match the paragraph divisions in the table. Also, paragraph section numbers need not be assigned in numerical order. For example, you may assign 1b, 2a, and 1c to sentences 1-3, respectively. See the coding of the sample letter in Appendix E for more examples. (Items 1-13 in the following table of paragraph section numbers is from Appendix 3 of (Dickerson, 2003), which is an adaptation of the outline given in (Baker et al., 1998). Item 14 was added for this study.)
<table>
<thead>
<tr>
<th>Topic</th>
<th>Subtopic</th>
<th>Paragraph Section Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>Date of visit and name of clinic</td>
<td>1a</td>
</tr>
<tr>
<td></td>
<td>Reason for referral and source of referral</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td>Purpose of letter</td>
<td>1c</td>
</tr>
<tr>
<td>Family history</td>
<td>Positive family history</td>
<td>2a</td>
</tr>
<tr>
<td></td>
<td>Negative family history</td>
<td>2b</td>
</tr>
<tr>
<td></td>
<td>Uncertain family history</td>
<td>2c</td>
</tr>
<tr>
<td>Medical history</td>
<td>Relevant illnesses/medical concerns</td>
<td>3a</td>
</tr>
<tr>
<td></td>
<td>Positive report of physical manifestations of condition</td>
<td>3b</td>
</tr>
<tr>
<td></td>
<td>Negative report of physical manifestations of condition</td>
<td>3c</td>
</tr>
<tr>
<td>Developmental history</td>
<td>Major milestones achieved at appropriate times</td>
<td>4a</td>
</tr>
<tr>
<td></td>
<td>Delayed milestone achievement</td>
<td>4b</td>
</tr>
<tr>
<td></td>
<td>Evaluation by developmental specialist performed</td>
<td>4c</td>
</tr>
<tr>
<td>Physical exam</td>
<td>Relevant clinical exam findings</td>
<td>5a</td>
</tr>
<tr>
<td></td>
<td>Condition suspected based on clinical exam</td>
<td>5b</td>
</tr>
<tr>
<td>Natural history</td>
<td>Clinical features of condition</td>
<td>6a</td>
</tr>
<tr>
<td></td>
<td>Prognosis/progression of condition</td>
<td>6b</td>
</tr>
<tr>
<td></td>
<td>Testing</td>
<td>6c</td>
</tr>
<tr>
<td>Inheritance</td>
<td>DNA – Genes – Chromosomes</td>
<td>7a</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------</td>
<td>----</td>
</tr>
<tr>
<td>1+</td>
<td>Dominant</td>
<td>7b</td>
</tr>
<tr>
<td></td>
<td>Recessive</td>
<td>7c</td>
</tr>
<tr>
<td></td>
<td>X-linked</td>
<td>7d</td>
</tr>
<tr>
<td>Risk assessment</td>
<td>Recurrence risk</td>
<td>8a</td>
</tr>
<tr>
<td>1</td>
<td>Risks for other family members</td>
<td>8b</td>
</tr>
<tr>
<td></td>
<td>Prenatal diagnosis available</td>
<td>8c</td>
</tr>
<tr>
<td>Test results</td>
<td>Positive results</td>
<td>9a</td>
</tr>
<tr>
<td>1</td>
<td>Negative results</td>
<td>9b</td>
</tr>
<tr>
<td></td>
<td>Ambiguous results</td>
<td>9c</td>
</tr>
<tr>
<td></td>
<td>Testing declined/postponed</td>
<td>9d</td>
</tr>
<tr>
<td>Recommendations</td>
<td>Medical management</td>
<td>10a</td>
</tr>
<tr>
<td>1+</td>
<td>Referral(s) to other departments/specialties</td>
<td>10b</td>
</tr>
<tr>
<td></td>
<td>Follow-up visit</td>
<td>10c</td>
</tr>
<tr>
<td>Psychosocial issues/concerns</td>
<td>Parental concerns</td>
<td>11a</td>
</tr>
<tr>
<td>1</td>
<td>Patient concerns</td>
<td>11b</td>
</tr>
<tr>
<td></td>
<td>Social issues</td>
<td>11c</td>
</tr>
<tr>
<td>Additional resources</td>
<td>National support group contact information</td>
<td>12a</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>0</td>
<td>Local support group contact information</td>
<td>12b</td>
</tr>
<tr>
<td>0</td>
<td>Patient literature</td>
<td>12c</td>
</tr>
<tr>
<td>0</td>
<td>Internet resources</td>
<td>12d</td>
</tr>
<tr>
<td>Closing</td>
<td>Return visit scheduled 4-6 months</td>
<td>13a</td>
</tr>
<tr>
<td>1</td>
<td>Return visit scheduled 1 year</td>
<td>13b</td>
</tr>
<tr>
<td>1</td>
<td>Contact clinic with questions</td>
<td>13c</td>
</tr>
<tr>
<td>Other</td>
<td>For politeness and/or to convey empathy</td>
<td>14a</td>
</tr>
<tr>
<td>0</td>
<td>Factual information not covered in 1-13 above</td>
<td>14b</td>
</tr>
</tbody>
</table>
Appendix B. Writing Pattern Tags

These tags are used to identify words or phrases that play a role in writing patterns. In this example using sentence (7) from (Baker et al., 2002), the tags added by the coder are shown in bold:

(7)  \(<\text{probability-7.1} \text{ Approximately 80\%} >\text{ of individuals affected with}\ \<\text{genotype-7/population NF} >\text{ have }\<\text{symptom-7.1/population mild to moderate symptoms}>; \<\text{probability-7.2} \text{ about 20\%} >\text{ have}\ \<\text{symptom-7.2/population more significant problems}>.\)

A tag encloses relevant parts of the sentence in angle brackets (<>). Tags may not be nested; in other words, you may not put one bracketed phrase inside of another bracketed phrase. Immediately following the lefthand angle bracket, each tag contains one or more descriptive keywords added by the coder, e.g., \(<\text{symptom-7.1/population}.\) The keywords that you may use are defined below. As in the example, a tag may contain several keywords separated by slashes (\(/\)). The first keyword following the lefthand bracket, e.g., \(<\text{symptom} \text{ in the above, is called the primary keyword} \text{. Primary keywords are numbered so that the tags can be referred to uniquely without confusion. Each primary keyword is numbered first with the sentence number; then if there is more than one tag with the same kind of keyword in the sentence, decimal numbers (.1, .2, .3, etc.) are added, e.g., \(<\text{symptom-7.1} \text{ and symptom-7.2} \text{ above}.\)

Following the primary keyword, \text{owner modifier keywords are given, e.g., population in the above example. Following the owner modifier keyword, the phrase that is being tagged begins, e.g. "mild" is the first word of the phrase tagged as symptom-7.1 above. Place the tag (left angle bracket and keywords) to the left of the first word of the phrase. (By phrase, we mean a unit of syntactic structure. Don't worry too much if you have never studied syntax seriously. The examples should be sufficient to get our point across.) Note that noun phrases often begin with a word whose grammatical category is an article ("a" or "the"), possessive pronoun ("his" or "her", etc.) or a quantifier (a word such as "some", "many", etc.). The righthand angle bracket is used to mark the end of the linguistic head of the phrase (linguistic head is a grammatical term -- for example, the head of a noun phrase is a noun, the head of a verb phrase is a verb, etc.). Sometimes it may be difficult at first to decide where to mark the end. Consider this coded excerpt from sentence (3) in (Baker et al., 2002):

(3)  … Philip was diagnosed as having \(<\text{symptom-3.1/proband a pseudoarthrosis} >\text{ of the left tibia and also was noted to have}\ \<\text{symptom-3.2/proband several cafe'-au-lait spots}> …\)

Note where the coder placed the right angle bracket for symptom-3.1; it was placed after "pseudoarthrosis", but why not after "tibia"? In this example, "pseudoarthrosis" is a head noun and "of the left tibia" is a grammatical modifier of the head noun (in other words, it provides additional information about the head noun). The reason for the policy of
placing right angle brackets after the head of the phrase is to make reading and coding easier; it will be understood that the description of the concept may include modifiers that immediately follow the head even though they are not inside of the brackets! To sum up, even though the \texttt{symptom-3.1} tag refers to the specific concept of \textit{pseudoarthrosis of the left tibia}, the right angle bracket is placed at the point in that text immediately following the head noun, "pseudoarthrosis".

Also, occasionally you will find a discontinuous sequence of words that seems to logically belong to the same tag; consider this example:

\ldots children with \texttt{symptom-7/population} severe to profound > recessively inherited \texttt{symptom-7-continued/population} hearing loss >

In the above, the words describing the symptom are interrupted with the phrase "recessively inherited". To handle this type of situation, you can use a \texttt{-continued} prefix. The meaning of the tagging convention in this case is that the full text of \texttt{symptom-7} is "severe to profound hearing loss". Lastly, occasionally it is clear from the context that a phrase refers to more than one individual who must be given distinct decimal numbers so that each can be shown as a separate node of the BN diagram (discussed in Appendix C). For example, the coded phrase, \texttt{genotype-29.1/mother//29.2/father NF >}, is used to signify that the phrase "NF" refers to each of the two parents, corresponding to nodes \texttt{genotype-29.1} and \texttt{genotype-29.2}. In cases where the multiple persons have differing owner modifiers you can use double slashes (//) to make clear which numbers are associated with which owner modifiers.

The \textbf{primary keywords} are defined as follows:

- \texttt{history}: demographic and predispositional factors such as gender, age of onset of symptoms, ethnic or geographic origin, family history of a disorder; \textit{history} covers factors that may not cause a genetic condition directly but may be associated with increased risk of having a genetic condition.
- \texttt{symptom}: observed traits such as hearing loss for which testing, diagnosis, and/or treatment could be sought and that may be the result of a genetic condition. (Observed traits are not limited to traits that can be detected visually.)
- \texttt{finding}: other observed traits noted by medical professionals that may be relevant to diagnosis but are not the chief complaint of the patient; the traits may or may not be the result of the same genetic condition as the symptoms. Also, findings differ from history in terms of causal order: a finding may be the result of a genotype but a history factor for a person cannot be the result of that person's genotype.
- \texttt{test}: genetic or other medical test
- \texttt{result}: result of test.
- \texttt{genotype}: a pair of alleles that may be the underlying cause of the symptom(s), the finding(s), and/or the test result(s). In some cases, a letter may discuss a mutation that results in differences in the same pair of alleles in the same person at different points in his/her lifecycle and/or in different parts of his/her body. In these cases, you should add one of the \textit{genotype qualifiers} (\texttt{zygote}, \texttt{germ cell}, or \texttt{somatic}) to each of his/her genotype tags to avoid ambiguity:
• **genotype (zygote):** the genotype of the person at conception. (Note: if no genotype qualifier is added, then genotype alone is assumed to signify this. In other words, you do not have to add this qualifier unless there would be ambiguity otherwise.)

• **genotype (germ cell):** the genotype of one or more of the person's germ cells

• **genotype (somatic):** the genotype of one or more of the person's somatic cells

• **mutation-event:** an event that has resulted in a difference in the same pair of alleles in the same person at different points in his/her lifecycle and/or in different parts of his/her body. For example, due to a mutation event, a person's genotype in (some or all of) his/her germ cells may be different from his/her genotype for that allele in the rest of his body (somatic), which is the same as the genotype that he/she was born with (zygote). This can be diagrammed as follows, with arrows showing causal influence and AA or Aa representing the pair of alleles:

![Mutation Event Diagram](image)

Another example is when a mutation event occurs in the somatic cells (leaving the germ cells unaffected) resulting in two different (somatic) genotypes for the same pair of alleles in different parts of the person's body:

![Mutation Event Diagram](image)

• **biochemistry:** a possible effect of a genotype on the biochemical makeup of the individual, e.g., shortened Connexin 26 (a protein)

• **physiology:** a possible effect on the physiological function of an individual caused by the individual's genotype or biochemistry. For example, insufficient production of insulin by the pancreas is a physiological function that could be caused by a protein abnormality (biochemical effect) that itself could have a genetic cause.

• **probability:** an expressed or implied belief in the (past, current, or future) possibility, frequency or probability of an event or condition. Although in some cases the writer's assessment of frequency or probability may be definite and is based on empirical studies or accepted scientific theory, the probability keyword also covers less definite or more subjective indicators. For example, numerical or qualitative phrases that could be used to describe frequency include "often", "3 out of 4", "75%", "many"; examples of phrases implicating possibility or probability are "may", "will", "would", "if", "whether", and "likely" and their negation (such as "would not"). In
cases where from context it is clear that the writer wants to convey a 100% or 0% chance from use of verb phrases (such as "are" or "are not", respectively), you may even tag them. Other phrases sometimes used to convey probability include "cause" and "due to". However, in some cases a sentence may express a probability statement with no explicit indicator of the degree of certainty, i.e., with no phrase that can be annotated with this tag. See section D for information on [IMPLIED] probability.

You should assign primary keywords based upon your judgment about the meaning of the text (instead of based on having the same parts-of-speech as the examples). Also, in a few cases, it is necessary to tag a phrase referring to a group whose members have more than one type of primary keyword with a hybrid primary keyword such as symptom-history; for more information, see section C2 on groups.

**Owner modifier** keywords are added to all primary keywords except the probability keyword, and are used to keep track of which individual a tagged phrase is discussing. In some cases, (A) the tagged phrase is about the proband and/or members of his/her pedigree ("family tree"), including deceased and future members; in other cases, (B) the tagged phrase is about a population in general or a member of a generic pedigree. First decide whether the phrase fits cases (A) or (B). For case (A), only assign one of the following pedigree modifiers:

- **proband**: the person who is the main subject of genetic testing, diagnosis, and/or treatment; normally this is the client or the child of a client. **All other pedigree modifiers in the rest of this list should be used to describe relationships of other persons relative to the proband; note that the described relatives may be living, deceased, or not yet existing (such as future siblings, mates, or offspring).**

- **mother, father, parent**: genetic mother, father, or parent of proband; use **parent** when gender is not specified and only one parent is referred to. When both parents are referred to but the letter does not specify which is mother or father, you should arbitrarily tag one variable as **father** and the other as **mother**. This is to avoid confusing two nodes marked as **parent** that actually refer to different parents. (Warning: in some cases although the writer speaks of "the parent", you can tell from context that the writer means each genetic parent, and thus two tags, one for each, should be assigned to the phrase.)

- **brother, sister, sibling**: genetic brother, sister or sibling of proband; use **sibling** when gender is not specified.

- **son, daughter, offspring**: genetic son, daughter, or offspring of proband; use **offspring** when gender is not specified.

- **wife, husband, mate**: the proband's reproductive mate: female, male, or gender-unspecified, respectively; use **mate** when gender is not specified and cannot be inferred from other information such as the gender of the proband.

- for other persons related to the proband, create compound tags using the above tags in the order from the proband up or down the family tree to the person being described adding the possessive indicator, apostrophe-"s"; for example, a paternal grandmother of the proband's grandchild would be tagged as **father's-mother**, or a future son of the proband's granddaughter (through the proband's daughter) would be tagged as **daughter's-daughter's-son**.
In a few cases, the above owner modifiers might not be sufficient to avoid ambiguity, e.g., where a letter discusses two different sisters of the proband. In such cases, add distinct numbers to disambiguate the owner modifier, e.g., sister-1 and sister-2.

For case (B), assign the population tag followed by a pedigree modifier for the person's place in a generic pedigree. You may use your interpretation of the relevance of the discussion about the general population to the proband's case to help you assign the pedigree modifiers. Also, you may have to carefully consider the rest of the letter to determine which person in a generic pedigree is under discussion at some point. For example, population/mother and population/father might be used to tag phrases about carriers in a general explanation of autosomal recessive inheritance (e.g. in a letter for a case where the actual proband's parents are suspected of being carriers). After you identified which phrases to tag as the parents then you could tag phrases about one of their offspring as population/proband. (Note that, since all pedigree modifiers are interpreted relative to a proband, you would not want to tag that offspring as population/offspring, which would signify a generic proband's offspring!) If more than one offspring is discussed, then pick the one who seems to be most relevant to the real proband as the population/proband and tag the others as population/sibling, since they would be siblings of the person tagged as population/proband. If you are not able to determine an individual's place in a generic pedigree, then (as a last resort) the population tag may be assigned alone, i.e. without any pedigree modifier.

Appendix C. Network Graphs

C1. Graph diagrams

A Bayesian Network (BN) is a graphical model whose nodes represent discrete variables and whose arcs represent dependencies of conditional probability between variables. A BN provides an intuitive way of depicting causal and associational relations in medical genetics. For example, Figure 1 is a BN showing the cause-effect relations between genotypes of the proband's parents and of the proband, and between their genotypes and test results, symptoms, and findings. This BN also depicts the association between history factors and increased risk of certain genotypes. You will construct one or two BNs based on the phrases in the letter that you tagged; one for discussion about the proband's pedigree and/or one for discussion about the general population. For example, Figure 1 depicts discussion about a specific proband and his/her family; Figure 2 depicts discussion only about the general population. The primary keywords used as tags, except for the probability tag, define the types of nodes that you may use in your diagram. (The role of the probability tag in the BN is described in Appendix D.)

When constructing your BN diagrams, you should follow these conventions:

- Draw a test as a diamond; draw the other primary tags except probability as rectangles.
- Label a node with the keywords (including numbers, qualifiers and modifiers) of the first tagged phrase in the letter that refers to the node, e.g., see Appendix E. Note that more than one tagged phrase in a letter may refer to the same node, even phrases with
different wording; use your judgment about the writer's meaning to decide whether a tagged phrase refers to a node that you have already added to the BN. For example, if one tagged phrase describes an individual's genotype as "has no normal copy" and another tagged phrase referring to the same individual's genotype (for the same gene) describes it as "has two abnormal copies", then both phrases refer to the same node (and, in fact, to the same value of that variable). Since a node in a BN stands for a variable that can take on different values, if the letter discusses different possible values of the same variable then you should represent that in the BN as a single node too. For example if a letter contains a phrase tagged genotype-5/sibling in a discussion about the probability that a future sibling will be heterozygous (e.g. genotype-5 = Aa) and another phrase tagged genotype-6/sibling in a discussion about the probability that the same future sibling will be homozygous (e.g. genotype-5 = aa), then the two tagged phrases would be represented by a single genotype node. If you decide that a tagged phrase does not refer to a new node, then add its number to some node's list of Coreferences, as shown in Figure 2. However, in cases where one tagged phrase refers to a more specific concept than another, you should represent the two phrases as different nodes. To give an example of this situation, "a genetic condition" may have been tagged as genotype-3/proband, and "VCF" may have been tagged as genotype-5/proband. In this case, since VCF is a more specific type of genetic condition, then each tag would be depicted by its own node.

- Do not include the actual words from the letter in your drawing, i.e., do not label a node as hearing loss or as symptom-3/proband hearing loss. (We broke this rule in Figure 1 to make it more readable. For a BN diagram following the convention that you should follow, see Figure 2.)

- Draw arcs (solid or dotted arrows) to show direct causal or associational relations between nodes, respectively, that are stated, presumed, or implied in the letter and that are allowed by the table below. (A writer may not explicitly discuss all links so you will have to exercise some judgement to identify links that are only presumed or implied by the writer.) Only draw arcs between directly related nodes. For example, if nodes N1 and N2 are causally linked and N2 and N3 are causally linked, then it is not necessary to draw a causal arc between N1 and N3, except in the rare case where there is both a indirect relation between N1 and N3 through N2, as well as a direct relation between N1 and N3. Only the following relationships are allowed:

<table>
<thead>
<tr>
<th>Node at beginning of arc</th>
<th>Relationship</th>
<th>Nodes allowed at end of arc</th>
</tr>
</thead>
<tbody>
<tr>
<td>history of person P</td>
<td>association</td>
<td>genotype of person P</td>
</tr>
<tr>
<td>genotype of person P for allele pair A (optionally with qualifier Q)</td>
<td>causation</td>
<td>genotype for A, biochemistry, physiology, symptom, or finding of person P; or genotype for A of offspring of P (Exceptions: no causal link allowed (1) from P's somatic genotype to P's offspring's genotype; or (2) from genotype (P,A,Q) to genotype (P,A,Q’) unless Q’ happened after Q)</td>
</tr>
<tr>
<td>biochemistry of person P</td>
<td>causation</td>
<td>physiology, symptom, or finding of person P</td>
</tr>
<tr>
<td>physiology of person P</td>
<td>causation</td>
<td>symptom or finding of person P</td>
</tr>
<tr>
<td>mutation-event of person P</td>
<td>causation</td>
<td>genotype of person P</td>
</tr>
</tbody>
</table>
• **Test** nodes, which are not considered to represent BN variables but which represent actions that enable a causal relation to hold between a genotype for a person P and a result of a test for person P, should be drawn as a diamond between the genotype and result. In other words, the test node is drawn as a diamond in the middle of a causal arc from the genotype to the test result.

• After analyzing remaining intertag/internode relationships (described in section C2), show relationships in your BN diagrams as follows:

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Depicted in BN diagram as</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal relation</td>
<td>solid arrow</td>
</tr>
<tr>
<td>Associational relation</td>
<td>dotted arrow</td>
</tr>
<tr>
<td>Test for genotype</td>
<td>diamond on arc from genotype to result</td>
</tr>
<tr>
<td>Coreference</td>
<td>list of corefences in node</td>
</tr>
<tr>
<td>Group-to-members</td>
<td>solid bar with circle (circle next to the group node)</td>
</tr>
<tr>
<td>Subtype</td>
<td>arrow with open head pointing to the superconcept</td>
</tr>
</tbody>
</table>

For examples, see the BN diagrams for the (Baker et al. 2002) letter in Appendix E.

**C2. Adding relationship annotations to coding table**

As you construct the graphs, add the following information about intertag/internode relationships to the third column of the coding table.

• **Coreference:** Two tags T₂ and T₁ are coreferential (coded as "T₂=T₁") if they refer to the same node in the same BN. The tags must have identical primary keywords, keyword qualifiers, and owner modifiers; the tag on the righthand side of "=" should be the lowest numbered tag referring to the node (in other words, the number that is used to label the node in the BN diagram). Put this coreference information in the row of the sentence with the higher number. For example, if a BN node labeled `symptom-3/proband` is referred to later in the letter by the phrase tagged `symptom-7/proband`, then add this to the annotations in the row for (7): `symptom-7/proband = symptom-3/proband`. Note that when you are finished, every tagged phrase except for probability phrases must either (1) appear as a node in a BN diagram, or (2) appear in the Coreferences list of a node.

• **Group:** In some cases a phrase may refer to a node representing the concept of a group, whose potential members are described by other phrases (in the same and/or other sentences before and/or after the phrase referring to the group). Consider our coding of a sentence from (Baker et al., 2002):

  … Philip was diagnosed as having `<symptom-3.1/proband` a pseudoarthrosis > of the left tibia and also was noted to have
  `<symptom-3.2/proband` several café‘-au-lait spots > both of which
  are `<symptom-3.3/proband` features > of …

In this example, `symptom-3.3/proband` refers to a group whose potential members include `symptom-3.1/proband` and `symptom-3.2/proband`. In the coding table, this is represented as `symptom-3.3/proband = {symptom-3.1/proband, symptom-3.2/proband, …}`. (Supposing that in the above example, the letter had said "either" instead of "both", then the group would still be coded as shown here.) Put the group
equation in the row of the coding table containing the phrase referring to the group; if
the members are referred to later in the letter, just go back to this row to complete the
group equation.) Note that a group may be a member of another group. Also, note
that in some cases, a group may contain members with different primary keywords,
e.g. a symptom and a finding. In that case, tag the group with a hybrid primary
keyword composed of the primary keywords of its members, e.g. symptom-finding.

• **Subtype:** One node is a subtype of another (coded as "T_{sub}<T_{super}"") if T_{sub} refers to a
node representing a concept that is a subtype of the concept represented by the node
referred to by T_{super}, where T_{sub} and T_{super} must refer to nodes of the same primary
keyword category, such as *genotype*, and with identical owner modifiers and
qualifiers, and T_{sub} and T_{super} are each labels of nodes, not coreferences to nodes. For
example, `<symptom-3.1/proband pseudoarthrosis>` is a subtype of `<symptom-5.5/proband congenital defects of the bones>`.
However, `<genotype-1/parent GBJ2>` and `<genotype-5/population/parent genetic disorder>` should not be coded
as subtypes since they have different owner modifiers and refer to different BN
graphs. If necessary, use your own background knowledge about subtype
relationships between concepts in clinical genetics to help identify subtype
relationships in the letters. Note that the node that is a subtype of the other node (the
more specific concept) should always be written on the left side of ",<",, and that you
should only encode the subtype relationship between two nodes once, instead of each
time there is a coreference to one of the nodes. Put this subtype information with the
coding for the sentence containing T_{sub}. Only describe direct subtypes -- in other
words, if A < B and B < C, then do not annotate A < C since that information is
implied by the other two relationships. Direct subtype relationships are depicted in
the BN diagrams using an arrow with an open arrowhead (with the arrowhead next to
the more general concept). (If a letter has A < B and B < C and a causal relation from
D to B, draw the causal arc directly from D to B.) Note that sometimes a node F is
directly related to another node G by both the member-group and subtype-supertype
relationships. In that case, draw both types of links from F to G. (If A < B and B is a
member of C, but A is only indirectly a member of C --i.e., through B--, you should
not annotate A as a member or subtype of C.)

• **Analog:** Two nodes are analogs (coded as "T_AÆT_N") if T_N refers to a node in the
proband's BN graph and T_A refers to a node representing the same concept for the
analogous cells of the analogous person in the BN graph for the population, and each
of T_A and T_N is the tag of a node, not a coreference to a node. Note that in cases
where the tags have genotype qualifiers, the nodes must have identical genotype
qualifiers too. Note that you should only encode the analog relationship between two
nodes once, instead of each time there is a coreference to one of the nodes. Put this
analog information in the row of the sentence containing T_A. For example, if a node
referred to by the phrase tagged as *genotype-5/population/parent* in sentence (5) is
an analog of the node referred to as *genotype-4/parent*, then add the following to the
annotations in the coding table row for (5):
*genotype-5(population/parent) ➔ genotype-4/parent*.
If the proband's BN has no concept T_N at the same level of
specificity as T_A, but it has a node T such that T_N would be a supertype or subtype of
T and T_A would be an analog of T_N, then the analog relationship should be coded as
"T_AÆsuper(T)" or "T_AÆsub(T)" respectively. For example, the population graph
might contain <genotype-1/population/proband genetic alteration> (i.e. T_A) but the proband graph might not contain an analogous node but instead might contain <genotype-2/proband GJB2> (i.e. T); then the analog relationship could be annotated using super(T) in place of T_N, i.e., as genotype-1/population/proband → super(genotype-2/proband), i.e., genotype-1/population/proband is an analog of a concept that is a supertype of genotype-2/proband.

Appendix D. Probability Statements

D.1 Encoding probability statements

In addition to depicting causal or associational links between variables (nodes), a Bayesian network defines a conditional probability table for each node; i.e., for a node Y with direct arcs coming into it from nodes X1 .. X_n, the table for Y gives conditional probabilities P(Y | X1, X2, …, X_n), for all combinations of values of Y and X1 .. X_n. For example, if the node symptom-1/proband has exactly one incoming arc, which comes from the node genotype-1/proband, then a probability table for

P(symptom-1/proband | genotype-1/proband)

might contain the following information:

<table>
<thead>
<tr>
<th>genotype-1 = AA</th>
<th>genotype-1 = Aa</th>
<th>genotype-1 = aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>symptom-1: deaf = yes</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>symptom-1: deaf = no</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

(The above values come from classical Mendelian theory. Other sources of probability values include epidemiological or clinical studies.) However, the letters that you are analyzing typically do not describe all cells in the probability tables and it is not necessary for you to construct probability tables. Instead, in the third column of the coding table, you should write any probability statements that are expressed in the corresponding sentence of the letter. For example, the following tagged sentence

<probability-4 Approximately 60%> of
<history-4/population childhood>
<symptom-4/population hearing loss> is
<genotype-4/population genetic>.

expresses a conditional probability statement that would be annotated as follows in row 4 of the coding table:

P(genotype-4/population | history-4/population, symptom-4/population) = "Approximately 60%"

In general, when encoding a probability statement, give the variable names (e.g. history-4/population) but not the variables' values (e.g. "childhood"); and give the words actually appearing in the letter that were tagged as probability (e.g. "Approximately 60%"). Use the variable numbering of the tags in the sentences containing the probability statement, not the variable numbering used to label nodes in the graph; i.e., you may use tags that corefer to previous tags. Note that probability statements do not always include conditional variables, i.e., now and then you may find one in the form P(Y) (sometimes referred to as the prior probability of Y).
When encoding conditional probability statements based on the wording of the letter, it is sometimes difficult to decide which words describe conditional variables (i.e. the variables that are shown to the right of the vertical bar "|"). For example, even if you decide that a sentence refers to a BN containing nodes X and Y, where X is conditionally dependent on Y (i.e. the BN has an arc from Y to X), the information in the sentence might be given in either of the two forms 'P( X | Y) = p1' or 'P( Y | X) = p2'. Note that there is a way (using Bayes theorem) to calculate one from the other, but the two probabilities are not the same (i.e. p1 is not necessarily equal to p2). Since these two forms are not mathematically equivalent, you should take care in deciding which form is expressed in the text. For instance, the sentence in the above example does not assert that someone with genotype-4 and history-4 has approximately a 60% chance of symptom-4 (hearing loss)! Therefore, it would be inaccurate to encode it as

\[ P(\text{symptom-4/population} | \text{history-4/population} , \text{genotype-4/population}) = \text{"Approximately 60%"} \]

As a rule of thumb, you should try to encode one probability statement for each use of the probability tag in the letter. However, sometimes even though a sentence does not contain a word or phrase that you tagged with the probability tag, it may convey an implicit probability judgment. For example, suppose that a letter states that the proband has been diagnosed as having NF; this implies that the writer (in representing the belief of the person who made the diagnosis) has the belief that there is some chance (a probability greater than 0) that the proband has NF. Another example is a section of the letter talking about the future or a hypothetical situation; in this case, the word or phrase conveying the probability might have been given in a preceding sentence. In cases where there is not a tagged probability phrase, you may encode a probability statement if in your judgment it is consistent with the author's meaning. Just give the probability as "[IMPLIED]", e.g., \[ P(\text{genotype-3/population} | \text{symptom-4/population}) = \text{[IMPLIED]} \].

In standard probability notation, the items separated by commas on the righthand side of the vertical bar are interpreted as a conjunction. Do not employ commas on the lefthand side of the vertical bar. In a few cases, e.g. see sentence 25 in Appendix E, you will need to describe the probability that either or both of two events Y_1 and Y_2 occur (with or without conditions to the right of the bar); this is written as \( P(Y_1 + Y_2) \). However, in order to save space, you may use the following non-standard abbreviation. If a sentence contains more than one tag of the same type that would create multiple probability statements differing only in what appears to the left of the vertical bar, e.g.,

\[ P(\text{symptom-5.1/population} | \text{genotype-5/population}) = \text{"50%"} \]
\[ P(\text{symptom-5.2/population} | \text{genotype-5/population}) = \text{"50%"} \]
\[ P(\text{symptom-5.3/population} | \text{genotype-5/population}) = \text{"50%"} \]

then they can be abbreviated as follows:

\[ P(\text{symptom-5.1/population or symptom-5.2/population or symptom-5.3/population} | \text{genotype-5/population}) = \text{"50%"} \]

However, keep in mind that our non-standard "or" notation is not equivalent in meaning to the '+' notation!
Appendix D2. Classification of probability statements

For each probability statement $P( Y | X_1, X_2, \ldots, X_n)$, that you have encoded, also add the following two types of classification annotations about it to the coding table.

- Classify it as **progressive** or **regressive**. A progressive statement is defined as one where the order of presentation in the letter (in the current or previous sentences) follows the order of the causal arrows between the corresponding nodes in the BN; a regressive statement is defined as one where the presentation order in the letter is the reverse of that. (Ignore history tags when assigning progressive or regressive.) For example, if the letter contains "his symptoms are probably due to an abnormal allele", the statement would be classified as regressive since "symptoms" appears before "abnormal allele" in the letter, but symptoms are caused by genotype. In cases where there is no $X_1, X_2, \ldots, X_n$, or where at least one condition $X_i$ precedes $Y$ in the graph and at least one other condition $X_j$ follows $Y$ in the graph, do not classify it as either progressive or regressive.

- Classify it as **prior**, **predictive** or **retrospective** using the following table:

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$X_1, X_2, \ldots, X_n$</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>anything</td>
<td>no conditions given</td>
<td>prior</td>
</tr>
<tr>
<td>genotype, mutation-event, biochemistry, or physiology of person Y</td>
<td>any combination of history, symptom, finding, result for person Y and/or any variables for descendents of person Y, except that the conditions may not consist only of history variables for descendents of Y</td>
<td>predictive</td>
</tr>
<tr>
<td>genotype of person Y</td>
<td>any combination of symptom, finding, result for descendents of person Y</td>
<td>predictive</td>
</tr>
<tr>
<td>history, result, symptom, or finding of person Y</td>
<td>any combination of genotype, mutation-event, biochemistry, or physiology of person Y and/or any variables for ancestors of Y</td>
<td>retrospective</td>
</tr>
<tr>
<td>any variable of person Y</td>
<td>any combination of variables for ancestors of person Y</td>
<td>retrospective</td>
</tr>
<tr>
<td>anything else</td>
<td>anything else</td>
<td>(none of the above)</td>
</tr>
</tbody>
</table>

Here's an example of how these two classifications would appear in the coding table: **predictive, regressive**: $P(\text{genotype-3/proband} | \text{symptom-3/proband}) = "50\%"$. 
Figures referred to in previous sections of manual:

Figure 1

```
Figures referred to in previous sections of manual:

**History/proband**
- Age: child

**Genotype/proband**
- Abnormal copies of gene GJB2: two

**Result/proband**
- Altered GJB2: yes

**Biochemistry/proband**
- Connexin 26: abnormal

**Physiology/proband**
- Chemical equilibrium: abnormal

---

**History/mother**
- Family history of deafness: no

**Genotype/mother**
- Abnormal copies of gene GJB2: one

**History/father**
- Family history of deafness: yes

**Genotype/father**
- Abnormal copies of gene GJB2: one

---

**Symptom/proband**
- Deafness: yes

**Symptom/mother**
- Deafness: no

**Symptom/father**
- Deafness: no

---

**Coreferences:**
- 7
- 26.1
- 26.2
- 5, 6.1, 14, 15, 16
- 7
```

Figure 2

```
Figures referred to in previous sections of manual:

**History/proband**
- Coreferences: none

**Genotype/proband**
- Coreferences: 5, 6.1, 14, 15, 16

**Symptom/proband**
- Coreferences: 7

---

**History/mother**
- Coreferences: 26.1

**Genotype/mother**
- Coreferences: 26.1

**History/father**
- Coreferences: 26.2

**Genotype/father**
- Coreferences: 26.2
```

---

**Coreferences:**
- 7
- 26.1
- 26.2
- 5, 6.1, 14, 15, 16
- 7