Mathematical Description of Biological Structures, Mechanisms, and States

H. Joel Jeffrey
Department of Computer Science,
Northern Illinois University, DeKalb, IL 60115
jeffrey@cs.niu.edu

Abstract—-A new method for mathematically describing cellular and molecular structures, mechanisms, and states is presented. A novel mathematical formulation of structure is developed, and new mathematical formulations of structural complexity and similarity are introduced that take into account differences in composition and structure at all levels of detail and apply equally to structures, mechanisms, and states. A recursive formula for calculation of structural similarity is derived. The methods and mathematical formulations apply equally to cases in which we have complete knowledge and to those in which we have only incomplete or partial information. The formalism and the mathematical similarity definition are the full generalization of sequence and sequence similarity. They enable the creation of repositories of formal multi-level structural descriptions of biological entities and new search capabilities, such as searching for processes or structures similar to a specified one, or with specified structural or compositional deviations.

Keywords- multi-level structure; mathematical models of structure; quantifying structural similarity

1. Introduction

DNA and protein sequence databases are of great value to biologists. They have this value because they are formal. A DNA or protein is specified by a string of characters on a 4- or 20-letter alphabet; a PDB entry is specified by a set of formal atom names and locations. Because specifications are formal, formal measures of sequence similarity are possible, and software such as BLAST is used routinely to find sequences similar to a query sequence.

By contrast, the great majority of descriptions of structure, whether of mechanisms or structures, are not formal. They are in ordinary technical language, in some cases augmented with graphical devices such as interaction diagrams and pictures of key molecular constituents. Because it is represented primarily in natural language, most molecular biological knowledge cannot be handled algorithmically. We can search on amino acid sequence similarity, but we cannot, e.g., query for "all proteins with structure similar to degree *d* to human hemoglobin in the R state," and get back proteins that have subunits similar in number, shape, and inter-relationships to those in R-state hemoglobin.

The situation is even more problematical with respect to biological situations and conditions. The customary concept of state and its formalization represents only a narrow subset of the intuition of biologically important facts or situations, namely those involving only the values of attributes of objects. The more general concept, for which we use the term *state of affairs*, involves processes, objects, and other component states of affairs, related in various ways. When we say, e.g., "the p53 molecule is phosphorylated," or "the DNA is damaged," we are identifying states of affairs.

This paper, an expansion of the work in [1], presents a new formalism for describing biological mechanisms, structures, and states of affairs that is the generalization of the concept of structure to the entire range of processes, structures, and states of affairs encountered in biology. The strings of letters representing DNA and protein sequences are special cases of the formalism. Using the formalism, new mathematical formulations of the concept of shape and structural similarity is developed, one that takes into account differences in composition and structure at all levels of detail and applies equally to structures, mechanisms, and inter-constituent relationships. Additionally, a new mathematical formulation of structural complexity is defined.

One goal of this work is the creation of databases of molecular biological mechanisms, structures, and states analogous to those we now have for genetic and protein sequences, and software systems using the new formulations and other algorithms operating on such multi-level structural knowledge bases.

2. Specification by constituents and relationships

We approach the problem of specifying structure of processes, objects, and states of affairs by considering the thing to be described as comprised of immediate constituents with specific attributes and inter-constituent relationships (temporal, spatial, or some other kind), and identifying the constituents and the relationships with formal names, as in mathematical logic. (Following standard practice in mathematical logic, an attribute is formally a one-place relationship, but here we will, for expository purposes, identify them separately.) We use the abstract term "entity" as a cover term for object, process, or state of affairs. Thus, an entity is specified by specifying its constituents and the relationships between them. The relationships are the formalization of the intuitive idea of structure. Each constituent is itself an entity, and therefore its structure can be elaborated with a second ES, and so on, continuing to any level of description desired. We term the approach *Entity Specification (ES)*.

An entity may be an object (structure), process (mechanism), or states of affairs (the generalization of state). A state of affairs is an entity whose constituents may be any set of objects, processes, and other states of affairs, with the necessary constituent attributes and inter-constituent relationships. States of affairs allow the formalization of complex situations, such as the fact that a p53 molecule is phosphorylated, that the DNA is damaged, the rate or change in rate of a reaction, a concentration of a molecular species, etc. An important case is location: location is an attribute of a process or object, represented formally as with any other attribute. Transport processes thus are processes represented via the same formalism as other processes.

Specifications of the entity (process, object, or state of affairs) are done by identifying the immediate constituents and the relationships and attributes that must be present for the item to conform to the definition, and specifying all constituents, properties, and relationships with formal names and values. For example, the customary high-level description of hemoglobin is that it has two immediate constituents, which are the two αß subunits, and the angle between subunits, which has one value in the R state and another in T. The subunits are further described as having two constituents α_1 and $ß_1$, α_2 and $ß_2$, respectively. Processes have processes and objects as constituents: the steps of the process and the elements involved in it. Constituents of states of affairs may be processes, objects, or other states of affairs. The core of an ES is thus a list of the immediate constituents and a list of the n-place relations among the constituents.

As the hemoglobin example illustrates, a set of Entity Specifications of an entity is formal and multi-level. ESs employ the same logical device used in mathematical logic: the use of formal names that are expanded by use of structured descriptions employing other formal names. Names of entities and relations are formal designators; a formal description of an entity gives further detail, i.e., its constituents and how they are structured.

2.1 Entity Specification

An *Entity Specification* (*ES*) consists of an ordered pair (N, D), where:

- N is the (formal) name of the entity including, optionally, a list of alternate names and/or a numerical ID.
- D is a set of *paradigms*, the major varieties or descriptions of the entity. DNA transcription has two major varieties, eukaroytic and prokayotic. In addition, if is often desirable to specify alternate descriptions due to the state of knowledge of the phenomenon: conjectures, possible alternative mechanisms, etc. The paradigms are the distinct descriptions of the entity.

Each paradigm of D is an ordered triple (C, R, E), where:

- $C = \{(C_i, T_i)\}$, in which C_i are the constituents and T_i is each constituent's classification, an element of the set {P, O, S}, representing "object,", "process," or "state."
- $R = \{r_j\}$ is the set of n-ary *relationships* that must hold between the named constituents. Any relationship may be included, not only those definable in terms of physical locations or quantities. Equations specifying quantitative relationships, including differential equations, are formal relationship names.
- The constituents and their relationships specify the structure of the entity. Additional information specifies particular instances of the entity, by identifying which actual "things" (processes, objects, and states of affairs) fill the roles named by the constituents. This information we term the *eligibilities* for the entity: E is a set of ordered triples (C_j, i, r), in which, for each C_j,
  - C_j is the constituent;
  - i is the name of the actual individual;
  - r is the rule, or condition, under which i takes the role of C_j in the entity N.

2.2 Processes

Processes are multi-step changes in objects and how they are configured, i.e., the relationships between them. In addition, processes may occur in many versions, i.e., combinations of the stages that are all ways of the specific process occurring.

To represent this concept formally, the $\{(C_{i,} T_i)\}$ for a process are:

1. Two constituents, specifying the before-state and after-state.
2. A subset identifying stages, i.e., constituents C_j in which T_j = P. Some stages may be accomplished via two or more alternatives; these alternatives are included in this subset.
3. A subset identifying the objects, i.e., T_j = O.
4. A subset identifying the versions of the process. Each of these version constituents is a state of affairs, i.e., T_k = S, and its constituents are the stages that comprise the version.

The relationships between stages specify those that happen sequentially, in parallel, overlapping, or in any other temporal relationships.

The stages are the steps of the process, and the states are the usual concepts of the before- and after-states of a process.

2.3 Objects

Objects have only object constituents, and in that sense are simpler than entities in general or processes; each constituent of an object is of Type O.

Objects provide clear illustration of multiple paradigms. For example, the large subunit of a ribosome is commonly described as having a roughly spherical main body and three lobes (i.e., with 3 constituents), but it is also described as comprised of two rRNA chains (5s, 23s) and a number of proteins. Fig. 3 shows the ESs of the eukaroytic ribosome and its constituents' sub-constituents.

2.4. States of Affairs

States of affairs are the most general kind of entity, since the constituents may be any object, process, or other state of affairs.

Since there are no restrictions on the constituents of a state of affairs, the general Entity Specification of Sec. 2.1 is the form of a state of affairs. This means that any entity is formally equivalent to a state of affairs.

2.5 Examples

The kinds of entities most directly of interest in biology are mechanisms and structures. We illustrate mechanism ESs with the ES of cell cycle arrest and gene transcription, in which the Constituents are Stages, Versions, and Elements, and the relationships are the constraints on which Stage must complete before initiation of the next one. The eukaryotic ribosome is used to illustrate structure ESs.

2.5.1 Cell cycle arrest

Fig. 1 shows an Entity Specification of the process of damaged DNA stopping the cell cycle. (Formal names similar to ordinary English phrases and sentences are used, with the notational device of square-brackets to indicate use of formal Element names in Stages.)

Stage 4, the general process of gene expression, in this process specifies production of p21, the Individual for Element "a protein" (shown in Fig. 1 in brackets). Thus, what actually occurs is the expression of p21, the phenomenon to be described.

N: [A damaged DNA molecule] stops the cell cycle in [a cell] {P1}
S_{result}: not Occur(S-phase, a cell)
Elements: a damaged DNA molecule: DNA molecule D
a cell: the cell with the damaged DNA
D:
Paradigm 1: eukaroytic cell
Stages:
1. [A damaged DNA molecule] activates [an ATM molecule]
S_{result}: [an activated ATM molecule]
Elements:
a. a damaged DNA molecule: damaged DNA molecule D
b. an ATM molecule: ATM molecule A
c. an activated ATM molecule: activated ATM molecule A
2. [An ATM molecule] phosphorylates [a p53 molecule]
S_{result}: [a phosphorylated p53 molecule]
Elements:
a. a p53 molecule: p53 molecule p53P
b. a phosporylated p53 molecule: phosphorylated p53 molecule p53P
Condition: only after Stage 1
3. [An activated p53 molecule] binds to [DNA] at the p21 coding site
S_{result}: [a phosphorylated p53 molecule] bound to [DNA] at the p21 coding site
Elements:
a. DNA: the DNA of the cell
Condition: only after Stage 2
4. [A cell] produces [molecules of a protein] from [a gene]
S_{result}: [a number molecules] of [a protein]
Elements:
a. a protein: p21
Condition: only after Stage 3
5. [A p21 molecule] inactivates [a cyclin E:cdk2 molecule]
S_{result}: inactivated [cyclin E:cdk2 molecule E2M]
Elements:
a. a cyclin E:cdk2 molecule: cyclin E:cdk2 molecule E2M
b. cyclin E:cdk2 molecule: cyclin E:cdk2 molecule E2M
Condition: only after Stage 4
Versions: 1. 1-2-3-4-5

Figure 1: ES of "Damaged DNA stops the cell cycle"

2.5.2 Gene Transcription

Step 4 in Fig. 1 is gene transcription. Fig, 2 shows, in outline form, the information to be specified in an ES elaboration of it: 7 ESs, at 4 levels of specificity, with a total of 25 processes identified. The top, or overall, level is "a cell produces molecules of a protein from a gene," with three stages, which are its process constituents; Stage 3, with formal name "Ribosomes translate an mRNA transcript of a gene to molecules of the protein," is of course gene transcription, with 4 stages.

Cell cycle arrest and gene transcription illustrate a central feature of ESs and ES methodology, namely the multi-level logical structure of ESs. Any single ES presents a full (formal) specification of the process, object, or state of affairs *at that level of detail*. Further detail is specified by further ESs.

The hierarchical specification technique is the formal analog of the commonly used informal method for describing complex biological processes, namely a hierarchical description elaborating the process structure in finer and finer detail, beginning with a high-level description in terms of a small set of large "steps" and continuing with division of steps into sub-steps, etc.. The outline form of gene expression in Fig. 2 is an example of just such a hierarchical description.

Process: A cell produces molecules of protein from a gene
1. The cell transcribes the gene to an mRNA molecule
2. The mRNA transcript moves to a ribosome in the cytosol
3. Ribosomes translate an mRNA transcript of a gene to molecules of the protein
   3.1. A ribosome initiates translation of the mRNA transcript
      3.1.1 The small ribosomal subunit binds to the mRNA transcript near the start codon
      3.1.2 The small ribosomal subunit moves to the start codon
      3.1.3 A tRNA molecule binds to the start codon on the mRNA transcript
      3.1.4 The large ribosomal subunit arrives at the transcription site.
   3.2. The ribosome adds an amino acid to the peptide chain
      3.2.1 The amino acid on the aminoacyl tRNA molecule binds to the A-site on the large ribosomal subunit
         3.2.1.1 The first nucleotide of the tRNA anticodon forms a hydrogen bond with the first nucleotide of the codon
         3.2.1.2 The second nucleotide of the tRNA anticodon forms a hydrogen bond with the second nucleotide of the codon
         3.2.1.3 The third nucleotide of the tRNA anticodon forms a hydrogen bond with the third nucleotide of the codon
      3.2.2 The large ribosomal subunit joins the amino acid on the P-site with amino acid on the A-site
         3.2.2.1 Peptidyl transferase breaks the bond between the amino acid and the tRNA molecule at the P-site
         3.2.2.2 The P-site amino acid and the A-site amino acid form a peptide bond
         3.2.2.3 The P-site tRNA molecule moves to the E-site on the ribosome
         3.2.2.4 The A-site tRNA molecule moves to the P-site on the ribosome
      3.2.3 The P-site and A-site tRNA molecules move three nucleotides in the 3' direction on the mRNA transcript
      3.2.4 The E-site the releases the tRNA molecule attached to it
   3.3. The ribosome terminates the polypeptide chain.
      3.3.1 The release factor binds to the large ribosomal subunit
      3.3.2. A peptidyl transferase molecule catalyzes the addition of a water molecule to the peptidyl tRNA, breaking the bond holding the polypeptide to the tRNA
   3.4 The large and small subunits of the ribosome dissociate

Figure 2: Process of gene expression (outline form)

Complexity in a process is often due to complex relationships between stages, such as the initiation of one stage only upon completion of another stage, perhaps of an entirely distinct process. Further, often these conditions involve additional factors of several kinds, e.g., a combination of a concentration of a biochemical and the physical location of a ligand. In these cases the formal expressive power of Entity Specification, in particular the formal inclusion of any relationship between constituents, as in mathematical logic, provides the capability of capturing the actual condition.

A different and important source of complexity in biological processes is the common situation in which the "output" of one process is an input to another. Fig. 1 shows an example: Stage 4 specifies the general process of gene expression; the particular instance of this general process is the production of p21, the key molecule in Stage 5. Fig. 1 thus illustrates two additional aspects of Entity Specification as applied to complex processes: 1) specification of a general process instantiated to produce a specific result – in this case, a p21 molecule, and 2) formally specifying the requirement of the presence of an actual object for the process to continue. Thus, Entity Specification represents formally what one can say informally: "the p21 gene is expressed, producing a p21 molecule, which inactivates the cyclin E cdk2 molecule, and so the cell cycle cannot continue."

2.5.3 Ribosome structure

We illustrate object Entity Specifications with an ES of the ribosome, formalizing the customary description into RNA subunits and proteins, and their constituents in turn, as found, for example, in [11].

N: eukaroytic ribosome
Relationships:
a. molecular weight(ribosome.eukaroytic) = 4,200,000

D:
Paradigm: 1
Sub-objects:
1. [SRSU]
2. [LRSU]

Relationships:
a. molecular weight(SRSU) = 1,400,000
b. molecular weight(LRSU) = 2,800,000
c. adjacent(LRSU, SRSU)

N: LRSU
Relationships:
a. molecular weight(LRSU) = 2,800,000

Paradigm: 1
Sub-objects:
1. [5S RNA]
2. [28S RNA]
3. [5.8S RNA]
4. [Protein1]
…
52. [Protein49]

N: 5S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
120. [Nucleotide120]

N: 28S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
4700.[Nucleotide4700]

N: 5.8S RNA
Paradigm: 1

Sub-objects:
1. [Nucleotide1]
…
160. [Nucleotide160]

N: SRSU
Relationships:
a. molecular weight(LRSU) = 1,400,000
Paradigm: 1
Sub-objects:
1. [18S RNA]
2. [Protein1]
…
34. [Protein33]

N: 18S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
900. [Nucleotide1900]

Figure 3: ESs of the eukaroytic ribosome and constituents

This example illustrates two significant aspects of the use of ESs. First, just as with ordinary-English descriptions, completeness is not necessary. ESs may be used to represent as much as is known of the structure of the entity, or as much as is desired for the purpose at hand. The depiction of the ribosomes in [11] includes only the constituents and their weights, and one relationship, namely that the SRSU and LRSU are adjacent; this is the information formalized in Fig. 3. The 5S RNA constituent is described further only by noting that it contains 120 nucleotides, without specifying them, and this formalized by the subobjects Nucleotide1…Nucleotide120 above. Other descriptions of the 5S rRNA constituent of the LRSU of the eukaroytic ribosome specify the particular nucleotides; these are formalized by specifying the nucleotides (A, C, G, U) and their structure with ESs of the constituents of each: the nitrogenous base, the 5-carbon sugar, the phosphate groups, and the positional relationships between them.

Second, it illustrates that there is no single "correct" Specification of an Entity, represented by having multiple Paradigms. This is not a deficit of the ES approach, but is rather a formal representation of the fact that there are often multiple descriptions of the same thing. Thus, we find the LRSU described in terms of the 5S, 28S, and 5.8S rRNA constituents, and we also find it described in terms of constituents of the main body, central protuberance, ridge, stalk, and valley.

Paradigms are the means of formalizing multiple descriptions of the same thing. This is not simply a semantic technicality. In the next section, we will show how to use ESs to mathematically quantify the concepts of complexity and structural differences, and the definitions and algorithms are based on the constituents and relationships *in a description*, i.e., a paradigm. To put it differently, there is no such thing as the "real structure" of an entity – structure, mechanism, or state of affairs. Rather, there are multiple descriptions of the entity, in ordinary English or formal ESs, and it is only meaningful to compare descriptions of entities.

This however does not preclude the discovery of a canonical form of ES, or adoption of standards or conventions for creation of ESs. It may, for example, be desirable to adopt conventions for automatically and uniformly converting protein databases to multi-level ESs representing their secondary, tertiary, and quaternary structure.

2.5.4 DNA and protein sequence databases

In Section 1 we noted that the string representations of DNA and protein sequences are special cases of ESs. In DNA or RNA sequences, the letters "A," "C," "G," and "T" (or "U") are the constituents, and the single relationship is "adjacent." In a protein sequence, the constituents are the letters denoting the amino acids.

In actual sequences, the nucleotides (or amino acids) have relative positions specified by two angles and a distance, and in certain cases the more complete symbolic representation of the sequence including is useful: N_1 (φ_1,θ_1,d_1) N_2 (φ_2,θ_2,d_2) N_3 …. In ES representation of this kind of sequence, the letters are the constituents and the relationships are the three $\varphi(x,y)$, $\theta(x,y)$, and $d(x,y)$.

2.6 Algorithms

Any set of complete descriptions of processes and objects is suitable as the basis of software to analyze and retrieve information about them. When we have complete descriptions, it is relatively straightforward to construct algorithms that answer questions such as:

- How does process P take place, in these conditions?
  - Identify the version that satisfies all the necessary relationships r_i that must be satisfied for the constituent stages to take place
  - Identify the specific individuals that serve as each object.
- What happens if process P does not take place?
  - Find all processes Q in which there is a relationship r_i stating that stage Z of Q can occur only if P has occurred.
- What happens if there are none of object O (such as with knockout experiments)?

- Find all processes P in which O is an individual for element E in stage Z. Since no O is available, Z cannot occur, so all versions of P including Z cannot occur, and if there is no version of P without Z, P itself cannot occur.

These algorithms were successfully implemented and tested in [3].

Things are much more difficult when there is partial information at multiple levels. Many, perhaps most, molecular and cellular processes and structures are not fully understood down to the individual molecule level. This requires formal specifications integrating descriptions (knowledge) at multiple levels, and algorithms designed to operate on incomplete specifications at multiple levels. ESs appear to be the first formalization designed for this multiple-level representation task. Several software systems implementing algorithms for the above queries, and others, have been built based ES knowledge bases [2, 3].

3. Measuring Similarity and Complexity

Biologists routinely use concepts of complexity and similarity of structures and processes in analyzing situations, looking for related structures and processes, and formulating research questions. However, these concepts have, until now, only been articulated in an intuitive, rather than a formal, way and as a result researchers have not been able to use them directly. For example, it would seem obviously valuable to be able to query a database for enzymes similar to a given one, similar at all levels of structure. Retrieval by similarity – the ability to do BLAST searches – is the heart of the value of DNA and protein databases, but such searches are limited to primary sequence similarity.

In this section we use the ES formulation to mathematically define the concepts of complexity and similarity of any two entities. This makes possible the quantification of similarity between structures, processes, or states of affairs that reflects structural differences at every level, not only primary sequence similarity.

We first define the *structural complexity* of an Entity A, with N constituents A_1, ...,A_N and K relationships, recursively as:

$$SC(A) = \sqrt{N^2 + K^2 + \varepsilon \cdot \sum_{i=1}^{N} SC(A_i)^2} \qquad (1)$$

ε is an experimentally-determined multiplier modulating the impact of complexity of constituents, sub-constituents, etc.

In formally defining a similarity measure on pairs of arbitrary entities, such as biological structures or constituents of them, we want to take into account the following intuitions:

- The measure should be responsive to differences in attributes of the entities themselves.
- The measure should be responsive to similarity of structure. Structure differences in an entity are represented by having different relationships among constituents, or in having relationships to a different degree.
- When structure of the constituents of the entities is known, the similarity between A and B should reflect the similarity of the their respective constituents.

Accordingly, we define the structural distance between two entities in terms of the difference of (1) the properties of the constituents, and (2) how much the relationships between the constituents differ, as follows:

Assume we have two entities A and B whose structural similarity is to be calculated. Denote the constituents of A and B by A_1, ..., A_{NA} and B_1, …, B_{NB}, respectively. Let the properties of A and B of interest be p_1,…,p_M. Denote the relationships between A-constituents by r_1, ..., $r_{K,}$ and those between B-constituents by r_{K+1}, ..., r_{K+L}.

First, re-order the constituents of A in order of decreasing complexity, as measured by Formula (1), and similarly with the constituents of B.

We represent the properties of A- and B-constituents in a Property Matrix P, and the relationships between constituents with a Relationship Matrix R. P is defined as follows:

- P has M columns (one for each property of interest).
- Let the top NA rows of P represent the constituents of A, in order, and the next NB rows represent the constituents of B, in order.
- The matrix entries are the values of each constituent on each property p_i.
- If a constituent does not have property p_i, that matrix entry is blank.

	p_1	...	p_M
A_1			
...			
A_{NA}			
B_1			
...			
B_{NB}			

Figure 4: The Property Matrix P

P now represents the properties of interest of the A- and B-constituents. In order to meaningfully compare numerical values representing disparate properties, the value of P must be normalized. Accordingly,

- If any column has a value < 0, re-scale the values of the column by adding the absolute value of the minimum value of the column to each value in it. This makes the minimum value of each column 0.
- Normalize the values of P to the range 1 to 10, by setting
  $p_i(A_j) = 10 * (p_i(A_j) + 1) / (pmax_i+1)$,
  where $pmax_i$ is the maximum value of column i. (The value of 10 is an empirically-determined, selected to emphasize the relative importance of property differences compared to the number of constituents.)
- Set each empty entry of P to 0.

The values of the property matrix P are now between 0 and 10, 0 indicating the component does not have the property of that column, and 1 being the minimum actual property value.

We can now define the p*roperty distance* between any A- and B-constituents, A_i and B_j, by using the Euclidean distance between the corresponding A- and B- rows of P:

$$PD(A, B) = \sqrt{\sum_{i=1}^{M} (p_k(A_i) - p_k(B_j))^2} \qquad (2)$$

Since the rows of P representing properties of A-constituents are sorted in order of most-complex-first, as are the rows of P representing properties of B-constituents, we have a consistent procedure for deciding which A-constituent and B-constituent to compare. For example, if we are calculating the structural similarity of the ribosomes of two species, the calculated value would differ significantly depending on whether the two large subunits and two small subunits are compared, rather than the large subunits being compared to the small, and the ordering ensures that the large are compared to the large, etc.

We now use a similar matrix technique to calculate similarity based on *structure*, rather than properties. Structure is specified by relationships between A- or B-constituents, each relationship r_j being represented by an ordered n-tuples. Each relationship has a specific value. For example, in R-state hemoglobin, the angle between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers is 15°. Thus, the relationship has the formal name "angle," and angle($\alpha_1\beta_1$, $\alpha_2\beta_2$) = 15.

Denoting the number of A-tuples by NAT, and the number of B-tuples by NBT, we define R as follows:

- R has K+L columns, one for each relationship.
- Each row of R represents one tuple of A- or B-constituents, so there are NAT+NBT rows.
- The matrix entries are the values of the relationships have on the tuples. For example, the entry for the matrix at the row ($\alpha_1\beta_1$, $\alpha_2\beta_2$), column "angle," is 15.
- If a tuple does not have relationship r_k, the corresponding entry of the matrix is blank.

	r_1	...	r_K	r_{K+1}	...	r_{K+L}
A-tuple$_1$						
...						
A-tuple$_{NAT}$						
B-tuple$_1$						
...						
B-tuple$_{NBT}$						

Figure 5: The Relationship Matrix R

The values of R must be normalized in order to be able to make meaningful calculations with the values, as were the values of P:

- If any column has a value < 0, re-scale the values of the column by adding the absolute value of the minimum value of the column to each value in it.
- Normalize the values of R to the range 1 to 10, by setting
  $r_i(A_j) = 10 * (r_i(A_j) + 1) / (rmax_i+1_)$,
  where $rmax_i$ is the maximum value of column i. (As with P, 10 is an empirically-determined value chosen to emphasize the relative importance of relationship differences compared to number of constituents.)
- Set each empty entry of R to 0.

The A-constituent and B-constituent rows of P are ordered, to ensure a consistent calculation procedure. It is necessary to have a consistent scheme

for calculating the Euclidean distance between rows of R as well, for much the same reason. Therefore, for any A-tuple ta_j, let $tb_{\kappa(j)}$ denotes the B-tuple closest to ta_j, using Euclidean distance, i.e., the B-tuple most similar to ta_j.

We can now define the total distance between two Entities A and B in terms of the property distance and the structural distance:

$$TD(A, B) = \sqrt{PD(A, B)^2 + SD(A, B)^2} \qquad (3)$$

The *structural distance* SD(A, B) is defined recursively, using the matrix R, as follows:

Let MC = max(NA, NB) and MT = max(NAT, NBT). Then if both A and B have Descriptions, i.e., specified constituents and relationships, we define the structural distance SD as

$$SD(A, B) = \sqrt{(NA-NB)^2 + \sum_{i=1}^{MC} PD(A_i, B_i)^2 + \sum_{j=1}^{MT} \sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2 + \delta \cdot \sum_{i=1}^{MC} SD(A_i, B_i)^2} \qquad (4)$$

If NA > NB, $PD(A_i, B_i) = PD(A_i, 0)$ for i > NA, and similarly if NB > MC.

If NAT > NBT, $r_i(tb_j) = 0$ for NBT < j <= NAT, and similarly if NBT > NAT.

If NA > NB, there is no B-constituent to for the A-constituent, so $SD(A_i, B_i) = SC(A_i,)$, for NB < i <= NA, and similarly if NB > NA.

If either A or B have no Description, SD(A, B) = 0.

δ is an experimentally-determined discount factor reflecting the relative importance of the distance between constituents of A and B. (As with ε, preliminary work indicates a value of approximately 0.7 for δ.)

Intuitively,

- $PD(A_i, B_i)$ measures similarity of properties of each pair of constituents.
- $\sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_j))^2$ measures how much the constituents of A and B differ on relationship r_i; and the sum $\sum_{j=1}^{MT} \sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2$ measures the total difference in structures A and B, as articulated by the relationships r_i between A- and B-constituents.

If A and B are the same except for differing only in names of constituents and relationships (mathematically, are isomorphic), TD(A, B) = 0.

As the properties of A and B, the number of their constituents, the properties of the constituents, the structure of A and B, and the substructures of A and B diverge, TD(A, B) increases.

3.1 Examples

We illustrate the calculation of SD with two examples: the simple structures H_20 and NH_3, and the more complex case of eukaryotic and prokaryotic ribosomes, which illustrates the recursive calculation and the application of the measure in the presence of incomplete information.

3.1.1. Structural Similarity of H_20 and NH_3

For the purposes of this example, we ignore $PD(H_20, NH_3)$, so $TD(H_20, NH_3) = SD(H_20, NH_3)$, i.e., we calculate similarity due solely to structural differences between H_20 and NH_3. We assume that the properties of interest are atomic mass and electronegativity, and the relationships of interest are distance D and bond angle α between the central atom and non-central ones. (This example illustrates the fact that TD may be considered a class of measures rather than a single one, for the particular similarity values will depend on the properties and relationships included in the calculation. Choice of properties and relationships depends on the particular application.)

We suppose that the member attributes of interest in this case are atomic mass and electronegativity of the constituents, which give the P and R matrices shown in Tables 1 and 2:

	Atomic mass	Electronegativity
O	16	3.44
Hw	1	2.2
Hw	1	2.2
N	14	2.04
Ha	1	2.2
Ha	1	2.2
Ha	1	2.2

Table 1: P matrix for H_20 and NH_3

	D	α
(O, H)	95.84	104.5
(O, H)	95.84	104.5
(N, H)	101.7	107.8
(N, H)	101.7	107.8
(N, H)	101.7	107.8

Table 2: R matrix for H_20 and NH_3

Normalizing P and R and re-ordering rows so that the pairs of most similar rows are adjacent results in Tables 3 and 4:

	Normalized D	Normalized α
(O, H)	9.4	9.7
(N, H)	10.0	10.0
(O, H)	9.4	9.7
(N, H)	10.0	10.0
(N, H)	10.0	10.0

Table 3: Normalized P for H_20 and NH_3

Normalized atomic mass	Normalized electro-negativity
10.0	10.0
8.8	5.9
0.6	6.4
0.6	6.4
0.6	6.4
0.6	6.4
0.6	6.4

Table 4: Normalized R for H_20 and NH_3

From Formula (3) above, $TD(H_20, NH_3) = \sqrt{1 + 18.25 + 201.8} = 14.87$.

Similar calculations with CO_2 yield Table 5:

	H_20	NH_3	CO_2
H_20	0	14.87	7.23
NH_3		0	19.91
CO_2			0

Table 5: TD of H_20, NH_3, and CO_2

3.1.2 Structural similarity of eukaryotic and prokayotic ribosomes

Section 2.5.3 shows an ES of the eukaryotic ribosome. The analogous ES of the prokaryotic ribosome, from [11], is:

N: prokayotic ribosome
Relationships:
a. molecular weight(ribosome.prokayotic) = 2.500,000

D:
Paradigm: 1
Sub-objects:
1. [SRSU]
2. [LRSU]

Relationships:
a. molecular weight(SRSU) = 900,000
b. molecular weight(LRSU) = 1,600,000
c. adjacent(LRSU, SRSU)

N: large ribosomal subunit
Relationships:
a. molecular weight(LRSU) = 1,600,000

Paradigm: 1
Sub-objects:
1. [5S RNA]
2. [23S RNA]
3. [Protein1]

…
36. [Protein34]

N: 5S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
120. [Nucleotide120]

N: 23S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
2900. [Nucleotide2900]

N: SRSU
Paradigm: 1
Relationships:
a. molecular weight(LRSU) = 900,000
Paradigm: 1
Sub-objects:
1. [16S RNA]
2. [Protein1]
…
22. [Protein21]

N: 18S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
1540. [Nucleotide1540]

Denoting the eu- and prokaryotic ribosomes Rib-eu and Rib-pro, from (3) we have

$$TD(Rib_{eu}, Rib_{pro}) = \sqrt{PD(\text{Rib-eu, Rib-pro})^2 + SD(\text{Rib-eu, Rib-pro})^2} =$$

$$\sqrt{4.05^2 + SD(\text{Rib-eu, Rib-pro})^2} =$$

$$\sqrt{16.38 + SD(\text{Rib-eu, Rib-pro})^2}$$

Calculating SD(Rib-eu, Rib-pro) from (4), we have MC = 2 and $\sum \sum (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2 = 0$ because the only constituent relationship is adjacency, which is true of both eukaryotic and prokaryotic ribosomes. Setting $\delta = 0.7$, we have SD(Rib-eu, Rib-pro) =

$$\sqrt{(2-2)^2 + \sum_{i=1}^{2} PD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{2} SD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2}$$

From the normalized P matrix, the term

$$\sum_{i=1}^{2} PD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2$$

$$= 4.3^2 + 3.6^2 = 31.1.$$

The term $\sum_{i=1}^{2} SD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2$

$$= SD((\text{LRSU-eu, LRSU-pro})^2 + SD(\text{SRSU-eu, SRSU-pro})^2$$

Again from (4), SD(LRSU-eu, LRSU-pro) =

$$\sqrt{(52-36)^2 + \sum_{i=1}^{52} PD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{52} SD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2}$$

The term $\sum PD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2 = 0$, because the ESs here (which are a formalization of the description in [11]) do not include properties of the constituents of the LRSU.

The term $\sum_{i=1}^{52} SD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2$

becomes $SD(28S, 23S)^2 + SD(5.8S, 5S)^2 + SD(5S, 0)^2 + 34*0 + 15*1$, because the proteins of the prokaryotic LRSU correspond to 34 of the 49 of the proteins in the eukaryotic LRSU and there is no further specification of those proteins. (Were there such specifications, as there would be with a specification of the ribosomes' structure down to the amino acid or atom level, these terms would not be 0.) The only specification of structure of the rRNA constituents of the LRSU are the numbers of nucleotides in them, so $SD(28S, 23S)^2 + SD(5.8S, 5S)^2 + SD(5S, 0)^2 = (4700-2900)^2 + (160-120)^2 + 120^2 = 3{,}256{,}000$, and SD(LRSU-eu, LRSU-pro) =1509.8.

Similarly, SD(SRSU-eu, SRSU-pro) =

$$\sqrt{(34-22^2 + \sum_{i=1}^{34} PD(\text{SRSU-eu}_i, \text{SRSU-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{34} SD(\text{SRSU-eu}_i, \text{SRSU-pro}_i)^2}$$

$$= \sqrt{64 + 0.7 * SD(18S, 16S)^2} = 360.1.$$

Thus, SD(Rib-eu, Rib-pro) =

$$\sqrt{0 + 31.1 + 0.7*(1509.8 + 360.1)} = 114.5, \text{ and}$$

$$TD(Rib_{eu}, Rib_{pro}) = \sqrt{16.38 + SD(\text{Rib-eu, Rib-pro})^2}$$

$$= \sqrt{16.38 + 114.5}$$

$$= 11.4$$

3.2 Discussion

It was noted above (Sec. 2.5.3) that there is no single correct description of an entity. Correspondingly, there is no single "correct" value of TD or SD. Rather, as we have seen in the ribosome example, the calculation depends on the particular properties and relationships chosen as the basis of the calculation, and on the information represented in the particular ESs used, which may reflect information either omitted or unknown. Thus, in use, a researcher first specifies the properties and relationships of interest in the particular investigation, and uses structural similarity search to find structures, mechanisms, etc. similar *in those terms*. For example, it may be of value to find structures with a similar number of constituents in similar positional relationships, without regard to net charge on the overall structure, or structures very similar in shape (as measured by similarity of angle and position relationships) but ignoring the properties of a particular constituent. Or, as in ribosome example, the question of interest may be, "How similar are enzymes A and B in high-level structure, ignoring the fine structure of the proteins in each?"

The work of building large knowledge bases of comprised of Entity Specification of biological knowledge is in its initial stages, and work on building tools to support the creation of ESs is also in its initial stages. Because there are many possible descriptions of an entity, creating ESs that are accurate formalizations of existing, informal, descriptions requires some expertise in biology. This means the work must be done by people with some training in biology, or in collaboration with them. Experience to date, however, indicates that producing good ESs does not require professional-level expertise.

4. Relationship to other work

Entity specifications are based on the "representation formats" of P. G. Ossorio, the Object Unit, Process Unit, and State of Affairs Unit [2]. The representations formats, especially the Process Unit, were the basis of several successful computer systems implementing the algorithms enumerated at the beginning of Sec. 2.6. These included a number of query systems [3] and LDS/UCC, a large system to actually carry out the processes specified [4]. The LDS/UCC system shows the applicability of ESs to simulation, especially when knowledge of the structures and processes is incomplete and at multiple levels of detail, as is commonly the case in biology.

Representation formats have a clear similarity to frames, but are a substantial refinement of the concept of frame. The most important distinction is that the constituents of ES are those that must be present by definition of the entity, whereas a frame is defined simply as "things commonly found together" [5], or as in Protégé [7], a related concept. (Interestingly, while clearly a refinement of frames, Ossorio's work predates the introduction of frames by several years [6].) Entity Specifications may be viewed as a rigorous version of frames, combined with the mathematical logic approach to inclusion of relationships.

Class hierarchies, i.e., ontologies, are the most common representation of biological knowledge. As we have seen, a set of ESs specifying an entity at multiple levels of detail is a hierarchy, and thus a set of ESs has a superficial resemblance to an ontology. However, the resemblance is only superficial, specifically in that both ontologies and ESs are hierarchically structured. Ontologies are designed to represent class membership and inheritance information; ESs are designed to represent structure. In an ontology, a child node represents a particular kind of the parent node, and members of the sub-class inherit attributes defined on the super-class; in ESs, a child node represents a constituent of the larger entity. Properties of entities and relationships between them are not inherited by their constituents. Thus, both the information represented and the fundamental concept denoted by the parent-child relationship in the node hierarchy are entirely different. While relational or attribute knowledge are often included in ontologies, the relations and attributes are any of interest, not those that define the items and its structure.

Frame-based systems such as Protege can be used to define an ontology, but a slot in a Protege frame is not the same as a constituent of an entity. While using frames, descriptions using Protege are nonetheless ontologies, and thus represent class hierarchies, not constituents and inter-constituent relationships.

Some ontologies, such as Gene Ontology [8], include a specific relationship, *part_of*, which specifies that one item is part of another. This is the

same concept as that of entity being a constituent of another. The difference between GO and ES is that while both provide a mechanism for specifying that one item is part of another, only ES provides mechanisms for specifying the other facts about the parts: how the parts are related (the set of n-ary relationships $\{r_j\}$) and eligibilities $\{(C_j, i, r)\}$ that specify rules for which actual thing *i* may serve in the role of each constituent C_j. For example, in both GO and ES we can specify formally that $\alpha_1\beta_1$ and $\alpha_2\beta_2$ are parts (constituents) of hemoglobin, but in ES we can also specify formally that angle between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ is 15°. Since structure is defined by the relationships $\{r_j\}$, this means ES provides the formal mechanism for specifying all aspects of structures and mechanisms, rather than the bare fact that one thing is a part of another.

The GO relationship *is_a* allows definition of class hierarchies; and *regulates, positively_regulates and negatively_regulates*, identify inter-process relationships. These are the only relationships in GO. Entity Specification incorporates the formal specification of any relationship, as in mathematical logic. GO relationships are therefore special cases or instances of ES relationships.

Peleg *et al* [9] integrates hierarchical process descriptions and participant-role logic, using organization workflow models combined with the Tambis [10] ontology to model biological processes. Tambis has the difficulties of any ontology: the only relationships that can be represented in it are those derivable from the pre-defined base relationships combined by subset/superset and "is part of."

Certain of the concepts in this paper were also presented in [12], in the context of applications in the social sciences.

References

[1] International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies, 2008 (BIOTECHNO '08) Publication Date: June 29 2008-July 5 2008, Bucharest, Romania, pp. 100-108. ISBN: 978-0-7695-3191-5, INSPEC Accession Number: 10091005, Digital Object Identifier: 10.1109/BIOTECHNO.2008.13. Current Version Published: 2008-07-15

[2] P. G. Ossorio, *"What Actually Happens" – The Representation of Real World Phenomena*, *Descriptive Psychology Press*. Ann Arbor, MI: Descriptive Psychology Press, 2005. Originally published by: University of South Carolina Press, Columbia, SC, 1978.

[3] H. J. Jeffrey and A. O. Putman, "MENTOR: replicating the functioning of an organization," in *Advances in Descriptive Psychology*, vol. III, K. E. Davis, Ed. Greenwich, CT: JAI Press, 1983.

[4] H. J. Jeffrey, T. Schmid, H. P. Zeiger, and A. O. Putman, "LDS/UCC: Intelligent Control of the Loan Documentation Process," *Proceedings of the Second International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems*, University of Tennessee Space Institute, Tullahoma, Tennessee, June 1989. ACM Press, 1989.

[5] M. Minsky, "A Framework for Representing Knowledge," *The Psychology of Computer Vision*, P. Winston, Ed. New York: McGraw-Hill, 1975.

[6] P. G. Ossorio, *State of Affairs Systems: Theory and Technique for Automatic Fact Analysis*. Rome, NY: Air Force Rome Air Development Center, RADC-TR-71-1021971, 1971.

[7] Online: http://www.protege.stanford.edu. Last accessed May 6, 2009.

[8] Online: http://www.geneontology.org. Last accessed May 6, 2009.

[9] M. Peleg, I. Yeh, and R. Altman, "Modeling biological processes using workflow and Petri Net models," *Bioinformatics* 8, No. 6, 2002: 825-837.

[10] P. G. Baker, C. A.Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, "An ontology for bioinformatics applications," *Bioinformatics* 15, 1999: 510-520.

[11] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Molecular Biology of the Cell*, 4th ed., New York: Garland Science (Taylor & Francis), 2002, pp. 343-347.

[12] H. J. Jeffrey, "High-Fidelity Mathematical Models of Social Systems," AGENT 2007 Conference on Complex Interaction and Social Emergence, November 15-17, 2007. (Sponsored by Northwestern University and Argonne National Laboratory, in association with the North American Association for Computational Social and Organizational Science.) http://agent2007.anl.gov/2007pdf/Agent%202007%20Proceedings.pdf. Last accessed May 1, 2009.