Chapter3: DNA Sequence Alignment in the Light of the Abductive Nature of Cladistic Hypothesis Generation.

> ... so wiederholt sich für das engere Problem der Homologiefeststellung dasselbe, was für die gesamte vergleichende Morphologie und die Systematik gilt, es setzte eine Periode intensivster wertvoller Arbeit ohne vorherige Klärung der Arbeitsprinzipien ein. (... so happens the same for the more narrow problem of the establishment of homology that applies for the entire comparative morphology and systematics, a period of most intensive, valuable work started without prior clarification of the working principles.) Remane (1952:31 Chapter 2: The term homology and the criteria of homology).

INTRODUCTION

The section quoted above from Remane (1952) has a modern ring when applying his words to DNA sequence data. A wealth of molecular characters has been acquired, but the debate as to how to use, or interpret, this information in the context of organismal relationships is unsettled. Shall we use maximum likelihood, neighbor joining, or parsimony (e.g., Hillis *et al.*, 1994)? Is transition/transversion weighting appropriate (Kluge, 1997a)? Within the cladistic paradigm, sequence alignment is a crucial step because it establishes some concept of ‘homology’ in one of several competing incarnations currently en vogue. Sequence alignment is performed by a number of different computer programs (e.g., CLUSTAL of Higgins & Sharp, 1988; MALIGN of Wheeler

& Gladstein, 1994) and, alternatively, also manually. Wheeler (1996) claimed to have developed an alignment-free phylogenetic analysis program. However, it is only alignment free in so far as alignment and phylogenetic reconstruction are implemented in a single computer program. The advantages and disadvantages of certain alignment procedures have been considered in the framework of minimizing gap cost (Waterman *et al.*, 1991), maximizing computational efficiency (Waterman *et al.*, 1991), and evolutionary relevance (Gatesy *et al.*, 1993). Alignment has not, however, been discussed explicitly and rigorously (*contra* Mindell, 1991) in the context of establishing primary homologies. The latter have also been termed 'putative' or 'weak homologies', 'positional correspondences', or 'topographical identities' (de Pinna, 1991; Brower & Schawaroch, 1996). These terms often indicate marginally to substantially different view points applied to the same observational qualities.

In the following the process of DNA sequence alignment is illuminated in a bottom up approach. This necessitates the introduction of some philosophical concepts on the nature of observation, as well as a limited discussion of the mode of inference employed in cladistics as a whole, because the observational phase needs to be framed relative to the entire process of phylogenetic inference. The homology concept serves as a crucial reference point, because it is one of the central, indeed probably the most important, tenets of phylogenetics. It will become clear upon closer inspection that most current practices in DNA sequence alignment can not be upheld. appropriate alternatives are suggested throughout the chapter and are summarized at the end.

MOLECULAR AND MORPHOLOGICAL CHARACTER EQUIVALENCE

An important question to start with is whether morphological and molecular data are equivalent and whether these data can, must, or must not be treated in a comparable fashion. Figure 3-1 illustrates the pathway by which data—morphological as well as molecular—are treated. Here the commonalities of the different data types in terms of observation and explanation are emphasized.

The observational phase

From an epistemological standpoint, the acquisition of basic observational knowledge can be characterized as a three-step psychological process: sense perception, perceptual belief formation, and classification (Audi, 1998). Using the example in Figure 3-1, the general aspects of observation are outlined.

Individual specimens comprise the source of all observations (Figure 3-1: abalone mollusk at top). Regardless of specimen preparation, original observations manifest themselves in the form of sensory perceptions. At its most basic level, sensory perceptions are registers of our surroundings. As such, one does not perceive individuals per se, but particular properties expressed by those individuals. This *modus operandi* of perception accounts for our practice of describing organisms in terms of 'character states.' In contrast, living and inanimate objects only possess characters, not states. It would be quite impossible to have, say, a character 'gills' separate from the individual subsidiary components that make up a 'branched gill' as opposed to a 'filamentous gill.' The latter is nothing more than a specifier or a predicate of the former. For the sake of historic consistency, character and state here are used here.

In our example (Figure 3-1), a scanning electron micrograph (SEM) of a radula from an abalone is used as a generic place holder for morphological data, and a fictitious read-out from an automated sequencer or the manual reading of a gel stands for sequence data. The direct or indirect sense perceptions lead immediately to perceptual beliefs; the properties observed are independent of the observer as opposed to being

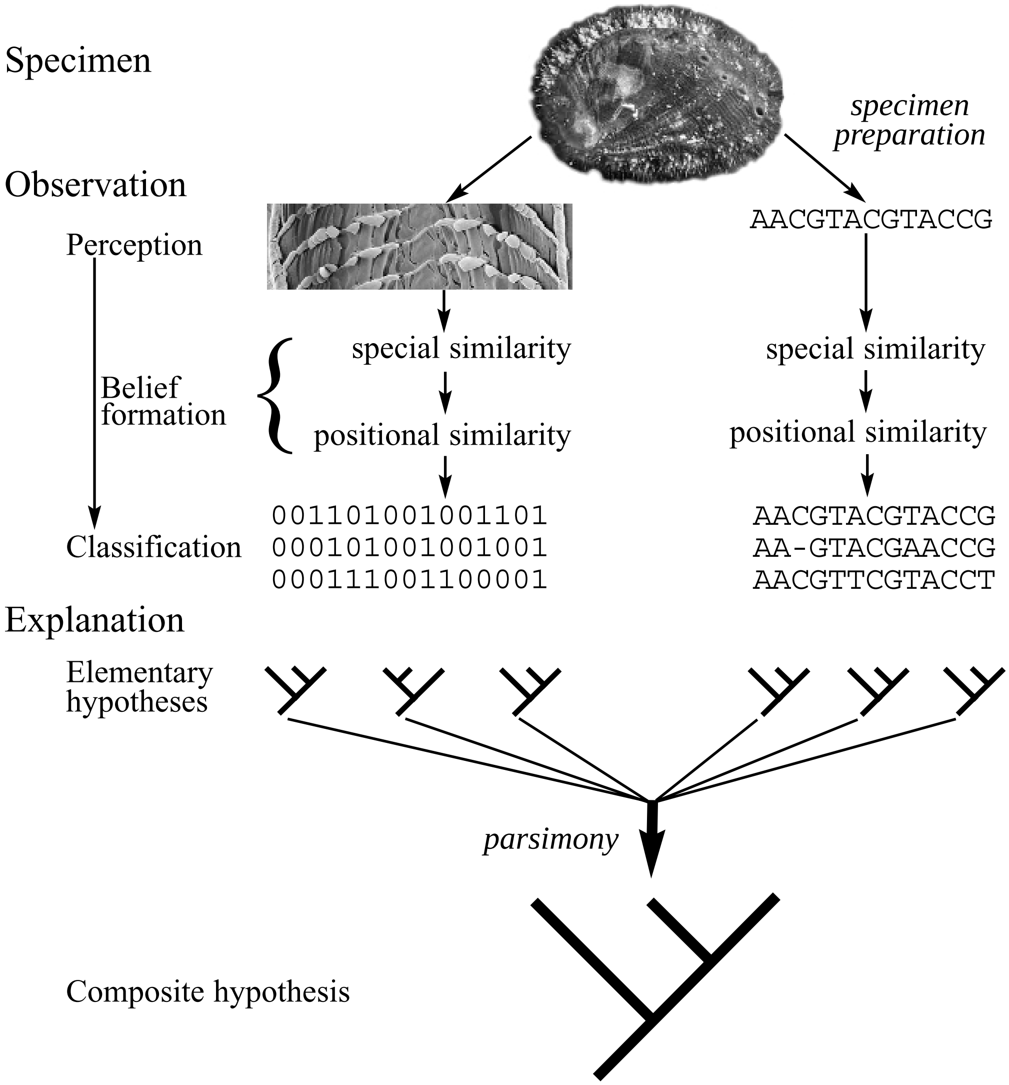

Figure 3-1. Treatment of morphological and molecular characters in cladistics with respect to homology. For details see main body of text.

hallucinatory events produced by the observer. It is on the basis of our perceptual beliefs that we claim particular properties, i.e., states, to exist. In other words, we engage in the process of naming those states by way of a predicate language. The act of naming properties observed of organisms is a base level act of classification. Here the term 'classification' is used not in the conventional systematic sense of ranking taxa, but in the sense of grouping observations. The step from perception to classification within the observational phase is the process of belief formation using auxiliary information (see below).

Are morphological and molecular observations equivalent as observations? One might make a distinction between the two on the following grounds. 1) Molecular data do not consist of transformation series. Bases can not show intermediate characteristics, for which reason they can not be treated like morphological data. The following, non-molecular examples, which show no intermediate character states, invalidate this argument: chromosome duplication, doubling of perianth, addition/loss of segments. 2) Ordering character states in sequence data is impossible. It has been argued that more explicit hypotheses of primary homologies are investigated, as the information contained in a fixed sequence of character-state transitions is higher than that in an unordered set of character states (Wise & Strong, 1997). The use of ordered characters may be viewed as a powerful tool, but it also harbors the potential for introducing notions on the evolution of character states that lack an empirical basis (Emberton, 1994; Sundberg & Hylbom, 1994; Wise & Strong, 1997). The above distinctions are viewed as unremarkable and not warranting a discrete morphological-molecular dichotomy. Consequently, all data as perceived shared similarities stemming from observations can and must be treated in an equivalent form, assuming our goal is the causal explanation of the phenomenon of shared similarity. As a corollary, commonalities of all observations need to be highlighted.

Note that the perception and the classification of a property do not convey a homology statement. The gray-scale values of each pixel in a SEM TIFF file is as little a primary homology statement as the bases in the original sequence read-out of any one taxon. The comparison of a set of SEM micrographs or a number of sequences in order to determine similarities still does not invoke primary homology, as no causal explanation is sought yet. To equate basic, named shared similarity (i.e., classificatory events) with homology is to remove any notion of causality. And, to remove causality is to remove any need for a term such as ‘homology’.

The explanatory phase

From the observed distribution of named states of any single character, an *elementary explanatory hypothesis* of homology is inferred. That homology hypotheses are explanatory resides in the Darwinian accounting of shared similarity by way of common ancestry, which incidentally, is nothing more than a replacement of Owen’s earlier causal accounting by way of archetypes. The very process of reasoning from some set of observed effects to an hypothesized cause, has typically been referred to as abductive inference, or “inference to the best explanation” (Harman, 1965; Fann, 1970). Unfortunately, the specifics of inference with respect to homology, as well as cladistics in general, have been at best poorly characterized. The importance here, however, is simply to point out that the inference of explanatory hypotheses is neither a deductive nor an inductive form of reasoning (Kluge, 1997b:92). This because the very structure of such hypotheses invokes unobserved (not necessarily unobservable) entities, in this case common ancestors, to causally account for what is observed, a particular character state distribution.

The claim that minimum-length cladograms are hypotheses with maximum explanatory power has been a fundamental tenet held by many advocates of parsimony. The abductive nature of homology hypotheses is not only consistent with this view, but in fact, the claim of explanatory power is completely dependent upon the abductive inference both from observed shared similarity to homology (= elementary hypothesis), and from homology to cladogram(s). In other words, cladograms are *composite explanatory hypotheses* based on the conjunction of all elementary hypotheses. Note that the (Darwinian) inference of homology is the association of common ancestry, i.e., cause, with shared similarity, or effects. The inference of cladograms from homology is nothing more than an additional abductive procedure wherein all elementary hypotheses are treated together applying the principle of common cause (cf. O'Hara, 1998). Abductive inference, or a logic of discovery (*sensu* Reichenbach, 1951), is used to infer first elementary and subsequently composite hypotheses, hence, nested abductive reasoning is employed (Josephson & Josephson, 1995; Richter, 1995; Fitzhugh, 1997; Fischer, 1997:375; Moser *et al.*, 1998). The question asked during the explanatory process is, 'why are the character states distributed as they are?' A primary homology hypothesis may be shown to be a 'confirmed', 'strong', or 'secondary homology' explained by common ancestry when considered in light of all shared similarity in need of explanation by way of common ancestry. Otherwise a case of homoplasy has been invoked (Farris, 1983). As an additional outcome of the generation of a composite explanatory hypothesis, one may even say a by-product, a pattern of relationships among taxa is revealed. One must be keenly aware that relationships among organisms are only a short hand form for the causal accounting of the character-state distributions among taxa. Unfortunately, it is the explanatory basis of cladistics, as well as homology, which seems to have been overlooked by proponents of methods other than parsimony.

Interestingly, the basis for the parsimony criterion in cladistics is not to minimize homoplasy, but to apply as consistently as possible the very causal notion of common ancestry used to infer homology. Any inference from effects to hypothetical cause(s) requires the association of those effects with some causal theory. To insist that homology, as well as cladistics, function in some theory-free, or model-neutral realm removes the ability to explain and, therefore, takes homology and cladistics out of the realm of science. The inference of a composite explanatory hypothesis requires an inference rule that is consistent with the major premise, or theory, used to infer elementary hypotheses. General evolutionary theory relative to inheritance and speciation justifies the use of parsimony as the criterion for hypothesis selection. This is nothing more than the application of the principle of the common cause (*e.g.*, Sober, 1988). To do otherwise would be to call into question the very homology hypotheses one has at hand, which would obviate cladistics altogether. If multiple sources of information are at hand, all of which require explanation by common ancestry, then they are combined in a 'total evidence' approach to obtain the (composite) hypothesis of relationship with the 'maximum explanatory power' for the character-state distributions in the form of a cladogram (Kluge, 1989; Kluge & Wolf, 1993). Unfortunately, the issue of 'total evidence', as well as 'explanatory power', have lacked any lucid philosophical underpinnings in the cladistics literature, except in the case of Kluge (1989) for the former. These matters will be treated elsewhere (Fitzhugh, in prep.).

To propose similarity perceptually, and to explain similarity with unobservables (i.e., ancestors), thus hypothesizing primary homology, are entirely separate inferential events, though, in practice, they are treated *as if* simultaneous to form a cohesive unit. An accepted similarity statement automatically opens the path to causally account for those similarities by way of common ancestry, as primary homologies. All observations entering an analysis must be treated in an equivalent manner, as all observations are subject to the same goal in cladistics: the causal explanation of shared similarity. Observations that do not satisfy the criteria of similarity (special similarity, positional correspondence: see below) are not included in the ensuing parsimony analysis. If the treatment of observation conflicts with the concept of homology, then these observations can not be included in a cladistic analysis unless the empirical basis can be established that one should not trust one's perceptual beliefs. On the other hand, to make such an assertion impinges not on cladistics, but on the very basis of inferring homology.

Cladistics as science

Cladistics does not follow the traditional Popperian concept of hypothetico-deductivism (*contra* Freudenstein, 1998:97). As no hypothesis is initially available, it can neither be confirmed, nor rejected, nor statistically tested. "Accordingly, there is no point in engaging in such fashionable academic probability games as musing about the probability of phylogenetic trees..." (Mahner & Bunge, 1997:48). Kluge (1997b), arguing that cladistics embodies Popperian tests, used the term 'test' in two different senses: A) provide reasons to select the most parsimonious tree against the hypothesis of a bush, which is not a Popperian test; B) evaluate an established phylogeny with new data, which again can not constitute a test since no test can rest solely on the very effects the hypothesis it is intended to explain. In contrast to Kluge's position, the actual evidence *for* so-called Popperian corroboration resides not in character-state distributions among terminal taxa, but must be those independent, subsidiary effects that must exist as a result of the specific initial causal conditions, i.e., the character states in the postulate ancestors, that lead to shared similarity. For the most part, the testing of historical explanations was a matter of little concern to Popper. The time spent with continued attempts to force Popper's views of science into history would be better served by investigating the philosophical studies of explanatory hypothesis testing. Either use of 'test' in cladistics does not constitute a deductive test as demanded by Popper. Interestingly, Kluge (1997b:92) also pointed out that "[h]ypotheses can never be proven true, as inductivists seek to do, nor proven false, as deductivists claim to be able to do." If cladistics employs neither induction nor deduction, indeed neither provides the inferential ability to generate explanatory hypotheses, then only abductive inference can be employed to address observations in need of causal explanation by which an hypothesis of relationship providing the maximum explanatory content is generated. As the ultimate goal of science is to causally account for observed phenomena (Popper, 1979:191; Moser *et al.*, 1998; Salmon, 1998), abductive, i.e., non-deductive, reasoning employed in cladistics is fully compatible with phylogenetic inference being a branch of science.

Figure 3-1 summarizes several of the points made above on the nature of cladistics, comparing explicitly the treatment of morphological and molecular characters. It is necessary to properly understand the entire sequence of actions during cladistic analysis, even if only a small part is treated here in detail. The following points are of particular concern for the discussion of sequence alignment presented here:

- Objects are perceived as sets of properties;
- All character states are comparable as observations;
- Homology is abductively inferred;
- Sequence alignment is the process establishing shared similarities (see below);
- Observation leads from perception to classification by way of perceptual belief formation (see below).

OBSERVATION: FROM PERCEPTION TO CLASSIFICATION

As noted earlier, observation can be separated into two distinct phases: sensory perception and classification connected by belief formation. Perception is applied in the sense of uninterpreted, sensory input, which has also been termed 'sensation' by Mahner & Bunge (1997). The matter is treated in a simplistic fashion here and do not address the question of whether perception/sensation itself is selective and, therefore, an interpretation of facts; see *e.g.*, Campenhausen (1993) and Josephson & Josephson (1995) for discussion. The two phases of perception and classification are linked by the mental process of belief formation (Audi, 1998). One readily recognizes a picture of a flower by comparison to one's mental library, but a computer would have difficulties to identify a flower from a graphics file. We arrive at the conclusion that this assembly of pixels represents a flower by forming a belief using auxiliary information (see below). Indeed, it has been recognized for some time that belief formation is itself an application of abduction (Devitt, 1997). By labeling entities with a name, which is a classification process, one proposes shared similarity. Similarly, with DNA sequences, the original read-out—the sensory input—must be converted using auxiliary information before the individual bases can be labeled shared similarities.

Sequence alignment is part of the process of observation leading from perception to classification. In that process gaps, which represent the absence of bases, are introduced in many instances. Nelson & Platnick (1981) have pointed out that the absence of a character state can not be observed. One does not note the absence of legs in snakes, but the continuity of the body wall. Corresponding arguments can be applied to gaps. What we can observe in DNA sequences is the adjacent position of two bases, or their homologues being set apart by inserted bases. Acknowledging Nelson & Platnick's (1981) point, the conventional use of 'absence' and 'gap' is continued here, despite being somewhat imprecise. As gaps are introduced during the observational phase, gaps can not be viewed as missing data, but are observations of absence of a base as the character state: gaps should be treated as a fifth character state. The same procedure has long been employed for morphological characters. Consider the legless condition of snakes as compared to other reptiles. If snake taxa were coded '?' for legs, then legs would be postulated as one of the possibilities during character state optimization, which is in conflict with the original observation. The 'legless' condition is a clear observation, if for no other reason than being a statement as to body form, which must be coded as such using a separate state. The same reasoning applies to gaps.

Determining character states and identifying the character to which the states belong are common to both morphological as well as molecular data. These two steps do not exist seperately, but are predicated upon one another. The two are part of an elementary classification system where various observations (= states) are united under the collective term of character. More formally, a character is nothing but a collective set of states (Goodwin, 1994). Note the necessary unity of the classification process: one can not identify a character without an underlying observation; and, what we observe are states, not characters. This process establishes shared similarity, which is open to explanation in terms of primary homology via common ancestry. Examples: A) character state known, position in question: this structure is yellow. In which position is this structure found? In the position of the petal (relative to other structures we identify *a priori* as flower-like properties). The petal is yellow. This DNA base is a G. In which position do I find it? It is in position 213. Base 213 is a G. Or, alternatively B) position known, character state in question: this is a petal. Which color does it have? This petal is yellow. This is position 213. Which base is found in this position? Position 213 is a G. The process of identifying states of particular characters is fundamentally one and the same process for any observation. Petals do not exist separate from the colors (and other properties) that allow for recognizing petals. Bases are only meaningful when classified in a particular position. ‘Organism X has a G’ is pointless, but ‘organism X has a G in position 213’ carries information.

Observed similarities leading to primary homologies are established using various additional information. Although the supplemental information used may depend on the type of data, the common theme is that auxiliary information is used for the identification of these similarities (see Remane, 1952). Two pathways can be identified.

i: Special similarity

Perceptual similarities are based on special, associated properties being shared. The associated properties are used to establish the similarity of the observations, which results in grouping states in characters. This special similarity argument is the most frequently used means for identifying similarities of morphological characters. Such an assessment is trivial with DNA sequences because the four bases are identical to their respective molecular structures.

ii: Positional correspondence

Remane (1952) used the agreement-in-position of a (morphological) character within the framework of the organism as the argument for similarity (see also Hawkins *et al.*, 1997). This is the only applicable argument that can be extended to DNA-sequence alignment (Hillis, 1994; Swofford *et al.*, 1996). Although the concept of positional correspondence was originally developed for morphological characters, its application to other types of characters, including DNA sequences, is straightforward (Brower & Schawaroch, 1996).

Sequence alignment

Three processes relevant to the context of sequence alignment can lead to differences between two sequences: mutation and insert/deletion (= indels). The goal of sequence alignment is to distinguish between these two sets of events in order to determine instances of shared similarity (cf. Kluge, 1997a:352). Four major methodological approaches can be considered. 1) Mutation or indels only. All evolutionary change is attributed to either of the mechanisms to the exclusion of the other. Such models are considered extremely unlikely and are not used. It is mentioned here for the sake of completeness. 2) Gap-cost function. A cost ratio for mutation to indel is applied and the least costly alignment is considered the most 'likely' (Li & Graur, 1991; Waterman *et al.*, 1991). 3) Parsimony-based. Parsimony is applied to find the alignment requiring the fewest number of character state transitions by employing a gap-cost ratio as well (Wheeler & Gladstein, 1994). 4) Manual alignment. The investigator matches the bases as well as possible by eye. The exact arguments for manual alignment can not be specified, but may well be a mixture of gap-cost function and parsimony. Much debate has centered on which alignment parameters are the most appropriate, because it is widely appreciated that different parameters yield different alignments, i.e., propose different shared similarity (*e.g.*, Gatesy *et al.*, 1993; Hillis, 1994; Bridge *et al.*, 1995; Wägele & Stanjek, 1995; Wheeler *et al.*, 1995; Wheeler, 1995).

One might prefer manual alignment as it forces one to consciously and critically evaluate every observation. Manual alignment also forces one to recognize areas of problematic alignment. An initial computer-facilitated alignment of any sort may reduce the manual labor, but as the critical work is done subsequent to the rough estimate, it becomes irrelevant which method of initial computer-facilitated alignment is chosen.

The pertinent issue is, however, the philosophical and epistemological bases for alignment as a similarity function.

OBSERVATION AND EXPLANATION

Homology is inherently comparative by the very fact that such statements are explanatory hypotheses, i.e., homology can not be applied to data from a single specimen, except in terms of individual ontogeny. Homonomy (= serial "homology") is not considered here and the discussion is restricted to interorganismal homology. The comparison of sequences, hence, sequence alignment, leads to elementary explanatory hypotheses in the form of primary homology statements. As Hawkins *et al.* (1997) have pointed out, the data matrix can be viewed as a set of assembled, primary homology statements.

Primary and secondary homology

Primary homologies are shared similarities causally accounted for through the postulate of a common ancestor at the level of each elementary hypothesis. At this level, homoplasy is an irrelevant issue. Secondary homologies are also shared similarities explained though a common ancestor, but in the context of a composite hypothesis. The latter, in the form of a cladogram, is not a simple summary statement of all elementary hypotheses, but a new hypothesis generated from all *relevant* evidence (cf. Wenzel, 1997). Relevance relates directly to the issue of total evidence, as only the *total relevant* evidence must be considered in non-deductive inference. Note that total evidence and relevance do not apply to deductive inference. Secondary homologies refer then to a restricted set of explanations, i.e., those that are explained by a common ancestor in the composite hypothesis. Homoplasies, however, despite being explanatory (Farris, 1983;

Siddall & Kluge, 1997:317), do not invoke a common ancestor at the point where they are shown to be homoplasious. Hence, the crucial point of homology is the explanation of shared character states by means of common ancestors at that particular level. The terms primary and secondary homology fulfill these conditions at their respective levels, hence, their use in phylogenetics is appropriate.

Separation of power

Observation, and explanation of observations, need to be independent of each other (Hawkins *et al.*, 1997). The observational phase must not dictate a preconceived explanation not evident in the observational process itself upon the observations. Otherwise the observational phase would interfere with the explanatory phase of the very observation. For morphological characters this problem has long been recognized in coding character states for a structure, such as 0: small, 1: large, 2: secondarily reduced. If the original observation 'small' is divided into two subcategories 'small' and 'secondarily reduced' despite no differences between the two conditions being discernible, then an explanation of the condition 'secondarily reduced' by way of reversal is included. Clearly, such coding practices are unacceptable and are based on preconceptions about the distribution of the data. The preconception of introducing an explanatory element into the observation may either originate outside the data (*e.g.*, following the "established" classification scheme), or may be derived from a cursory glance at the data matrix where some pattern is visually detected. Leglessness of some lizards and snakes must be coded the same, even though we infer from a wealth of other observations that it is a separately derived condition in the lizards. Otherwise we take the data beyond the observation of the absence of legs, and already include an explanation for the distribution of character states. (One may argue for the exclusion of this particular character, because it has already been explained, but this is a separate issue; see below.).

Confusion about the separation of observation and explanation in phylogenetics, including the necessary priority of observation over explanation, dates back to Huxley, and this confusion continues to this day (Brady, 1994). For instance, it has been argued that sequence data afford the unique possibility to combine the two steps—establishing shared similarities with the explanatory equivalents of primary homologies (= elementary hypotheses) and to explain them in a composite hypothesis—into one single procedure. In one case, sequence alignment is said to be abandoned altogether by a direct form of character state optimization (Wheeler, 1996; Wheeler & Hayashi, 1998). In a second approach, parsimony is used as the optimality criterion to align sequences (MALIGN: Wheeler & Gladstein, 1994) or to determine the sequence input order (Knight & Mindell, 1995). What is overlooked is that the establishment of primary homologies as elementary hypotheses is necessarily prior to the composite hypothesis, and the establishment of shared similarities must be independent for each character (see below), whereas the explanation of observed shared similarity at the level of the composite hypothesis is carried out in the framework of all data under consideration.

Mindell (1991) argued that, because the 'test' of congruence through parsimony minimizes homoplasy and respectively maximizes secondary homologies, the same maximizing methodology can also be applied to the establishment of primary homologies in the form of sequence alignment. Although the establishment of shared similarities and their subsequent explanation at the level of a composite hypothesis are contained in one logical sequence, the two steps must remain separate. The same criterion (parsimony) can not be used to find shared similarities, using as arguments for the determination of particular similarities all positions across all taxa of a particular sequence, and to causally explain the very same observations, using again all positions across all taxa of that particular sequence, because then circular reasoning ensues (Brady, 1994; Wheeler, 1995). This problem applies only if all data entering the observational phase are the sole bases for the explanatory process. This is the case with studies using a single, parsimony-aligned sequence to the exclusion of any other information, which is then analyzed using parsimony (*e.g.*, Wägele & Stanjek, 1995; Wheeler, 1995). Using parsimony to align sequences is comparable to the above mentioned cursory glance at the data matrix, only that the entire data matrix is relied upon for the establishment of shared similarities. The data matrix is the source of the bias. However, as soon as a single synapomorphy is added to the parsimony-aligned sequences, circular reasoning is no longer an issue. The parsimony argument at the explanatory level provides the best explanation for a more inclusive data set, not only for the parsimony-aligned sequence. However, the problems with character independence are still valid.

CHARACTER INDEPENDENCE

One of the central tenets in character coding is character independence. The term ‘independence’ is used here in the sense of potential of changes in character states not being under the influence of any other state; independence is distinguished from the existence of one state being predicated upon the existence of a second, the latter being better described as the problem of inapplicables. Separate characters as groupings of sets of states should only be considered if the set of states in each character has the potential for independent evolution (*e.g.*, Kluge, 1989; Brower & Schawaroch, 1996; Hawkins *et al.*, 1997; Luckow & Bruneau, 1997). The co-variation, or congruence, of multiple characters expected from common ancestry is not an indication for non-independence in the above sense, but is a matter of causal association by at least common ancestry. This distinction can also be characterized as the difference between an *interactive fork* (non-independence) and a *conjunctive fork* (co-variation due to common ancestry: Salmon 1998:296). As coded characters are independent, the inference of primary homologies for each individual character must be based on relevant auxiliary information (special similarity and positional correspondence: see above). Consider a problematic bone in the skull of some vertebrate. To establish its shared similarity we only consider features of the skull, but we do not take into consideration the hand or the rib cage. The latter are found beyond certain landmarks such as the non-arbitrarily identifiable neck, shoulders, and elbow. On the other hand, we can use position, structure, histology, etc., of a skull bone and its surrounding components to determine similarity, thus homology. In more abstract terms, the auxiliary information permissible to establish primary homologies is restricted in width, but potentially open in depth.

In molecular data the only available auxiliary information is positional, i.e., adjacent regions of the particular nucleotide provide the basis to postulate shared similarity. It is inappropriate to use auxiliary information beyond a secure anchoring point such as conserved regions of DNA. Therefore, characters beyond a conserved DNA region are ineligible to influence the establishment of the stretch within.

It is well-known that different genes are optimally aligned with different alignment parameters (gap weights: see below for further detail), so also various regions in one and the same gene may align “best” under different alignment conditions. To force one single set of alignment parameters on the entire sequence may deny optimal alignment to variable regions between conserved ones. From a process perspective, the same evolutionary conditions (*e.g.*, mutation rate, transition/transversion bias) are assumed to hold for the entire sequence. The parameters that produce an overall optimal alignment may be suboptimal for some regions, regardless of whether the minimization function is gap-cost or parsimony. In other words, process independence for each region is ill considered, resulting in a violation of character independence. The alignment process should pertain to the smallest fragments between adjacent conserved regions, which may be termed *minimal fragment alignment* (MFA). Alternatively, one might ask what is comparable to a global minimization with morphological characters. There is no comparable procedure.

Wheeler & Hayashi (1998) argued for using multiple gene sequences and even morphological data to help with sequence alignment. They pointed out that congruence is a well established decision making argument in phylogenetics, hence, the congruence between different data sets helps to establish similarities. The transition / transversion / gap ratio resulting in the best Mickewich-Farris metric between the data-sets was used. Congruence, indeed, has a sound grounding in cladistics, but at the level of the cladogram construction, i.e., the explanatory phase. The establishment of similarity, however, is carried out in the observational phase. Second, the issue of relevance arises. At the level of cladogram construction all character-state distributions with their similarities already identified are relevant. However, when the similarities are established only the appropriate auxiliary information is relevant. Clearly, morphological characters are entirely irrelevant when considering similarity issues at the level of DNA sequences.

How are conserved regions to be found? To identify them, indeed, a larger part of the sequence must be scrutinized, which still is not an argument to use global alignment for individual bases. We encounter the hierarchical nature in the relation of objects and their parts (part-whole relationship). We first need to identify the similarities of more inclusive structures (*e.g.*, gene, skull), then intermediate ones (*e.g.*, conserved region, neck), before detailed questions can be addressed (*e.g.*, base 213, bone A). The auxiliary information appropriate for each hierarchical level is chosen, keeping in mind the scope of the problem addressed. The auxiliary information deemed appropriate can not be determined with mathematical accuracy. However, some information is clearly inappropriate such as variable sites beyond a conserved region when aligning individual positons.

Mindell (1991) asserted that primary homologies established probabilistically through sequence alignment are of binary nature, i.e., qualitatively related so that structures are either considered to be homologous or not, but that none is labeled '87% homologous.' Mindell's rationale for the sudden transition from a probabilistic to a binary statement can not be followed, but it is taken at face value for the sake of the argument. The question arises of what would have to be done with morphological data to ensure equivalent treatment of all data? The morphological matrix would have to be aligned using parsimony (with the characters arranged anterio-posterior or dorso-ventral?). The interpretation of potentially resulting gaps would be challenging. As the approach is nonsensical for morphological data, the comparable treatment of all data is lost. Mindell's (1991) point is rejected.

OBJECTIVITY AND SUBJECTIVITY

Character selection and observation

Sets of molecular data are acquired by the use of certain pairs of primers. This choice of primers is a willful act, therefore, inherently subjective, as is the case with observation in general. The subjective choice is not limited to the stretch of DNA as an entity, but includes every single base. Character sets in need of explanation are chosen for morphology as well as in the case of genes or gene regions (cf. Salmon, 1998:306 on the relativity of relevant information). The choice of characters whose similarity is to be determined is related to the part-whole relationship of objects (Skull: dermatocranium: bone A. Gene: gene region: base 213). Features for which shared similarity is not clear (see below: Belief formation), hence, that can not reasonably be suspected to be similar between taxa, do not need to be explained. A character state distribution is already accounted for by virtue of recognized non-similarity, and consequently non-homology, and is no longer in need of explanation at the level of cladogram construction. The particular point is further explored when addressing the problem of character inclusion and exclusion below.

Alignment parameters

With molecular data, character states are given as a sequential reading of bases. The goal of alignment is the assignment of a limited number of shared character states to linearly arranged characters. The set of alignment parameters (gap weight with or without different extension costs: see Waterman *et al.*, 1991) is an assumption entering the analysis. These alignment parameters can also be viewed as the the quantifiers of an evolutionary model. As alignment parameters or the underlying evolutionary model can not be observed directly from DNA, they are extraneous to the data. Like any other weights (character weight, transition/transversion weight) they constitute assumptions, are inherently subjective, and generally are non-empirically justified. The advantage of computer-facilitated over manual alignment is the explicitly stated assumption in the form of specific alignment parameters (Gatesy *et al.*, 1993). The comparison to morphological observations will be taken up in the section ‘Belief formation’ below.

Objectivity seems to be some general goal in molecular phylogenetics (Messenger & McGuire, 1998:93) and is briefly discussed in Moore & Willmer (1996). It is claimed to be obtained in sequence alignment through a so-called sensitivity analysis (Wheeler, 1995). Hereby, a number of parameters are applied in order to investigate whether or when the alignment changes. If a wide range of parameters results in the same alignment, it is said to be more ‘robust’, giving results the guise of objectivity. If a change of parameters results in a different alignment then the operator is faced with the question of which parameters / alignment to choose. An important question in sensitivity analysis is at which interval and over which range the alignment parameters are used. Intervals may be subjectively chosen as integer numbers on a linear scale, or on a log scale with bases such as 2, e, or 10 (Wheeler, 1995). The range of values for gap weights has a logical lower limit of 0.5 because of the triangle inequality (Wheeler, 1993), but has an open upper bound. The only logical upper limit is infinity, i.e., change by mutation only. Most workers would argue that the infinity boundary is unrealistic. Restricting the range of alignment parameters used to less than the objective 0.5 to infinity automatically introduces subjectivity (Gatesy *et al.*, 1993). Whether the range is explicitly restricted in computer facilitated alignment or while performing manual alignment is irrelevant to the question of the subjective choice of parameters.

Manual editing

In many studies, a compute- generated alignment has been subsequently edited manually (*e.g.*, Collins *et al.*, 1994). Such a practice can not be justified in the context of objectivity. A single alteration of a computer-generated alignment instantaneously forfeits the advantage of the explicitly specified assumptions.

ALTERNATIVE CODING STRATEGIES

Unambiguous alignment is always unproblematic. The clearest case is with conserved regions harboring uninformative states. As soon as regions of unequal length are compared, we may face questionable or ambiguous alignment. Statements of shared similarity, thus primary homology hypotheses, may depend upon the alignment algorithm and parameters chosen. Even with a particular setting, multiple, equally parsimonious or costly alignments may be found, although not many programs report more than one. Ambiguous alignment can be addressed using various, flexible coding strategies. Each strategy is considered and the consequences for homology are illustrated with simple example.

This example consists of four hypothetical sequences (Figure 3-2: Original data). The sequences are characterized by initial and terminal conserved regions of five bases each, adjacent to variable regions spanning two positions, i.e., six and seven. In the variable region, taxon 1 shows a CC, and for taxon 2 a double gap (--) is found. Taxa 3 and 4 show one base each, C and A, respectively, and one gap. The absolute and relative positions of base and gap are unresolved for the last two taxa (Figure 3-2: Four possible alignments).

Two problems relating to homology will be encountered throughout Figure 3-2. 1) If any character state in one taxon has more than one homologous state in another taxon, i.e., the latter is coded in more than one column, then the test of conjunction is failed (de Pinna, 1991). This situation is indicated with underlined positions. 2) Contradictions in possible character state optimizations with original observations are flagged by the position *in italics*.

Elision

Finding a consensus of multiple alignments resulting from the choice of different alignment parameters has been addressed by Wheeler *et al.* (1995), who proposed a method called ‘elision.’ The alternative alignments for each taxon (Figure 3-2: Four possible alignments) are appended to one another (Figure 3-2: Elision) and analyzed

```
Original data          Four possible alignments
ATCTGCCACGTAC          ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC
ATCTGACGTAC            ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC
ATCTGCACGTAC           ATCTG-CACGTAC   ATCTGC-ACGTAC   ATCTG-CACGTAC   ATCTGC-ACGTAC
ATCTGAACGTAC           ATCTG-AACGTAC   ATCTG-AACGTAC   ATCTGA-ACGTAC   ATCTGA-ACGTAC

Elision                                                             Symplesiomorphies removed
ATCTGCCACGTACATCTGCCACGTACATCTGCCACGTACATCTGCCACGTAC                CCCCCCCC
ATCTG--ACGTACATCTG--ACGTACATCTG--ACGTACATCTG--ACGTAC                --------
ATCTG-CACGTACATCTGC-ACGTACATCTGC-ACGTACATCTG-CACGTAC                -CC-C--C
ATCTG-AACGTACATCTG-AACGTACATCTGA-ACGTACATCTGA-ACGTAC                -A-AA-A-

Case sensitive : a, c = A/T/G/C/-
Recoded:               Optimizations:
ATCTGCCACGTAC          ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC
ATCTG--ACGTAC          ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC
ATCTGc-ACGTAC          ATCTG--ACGTAC   ATCTGC-ACGTAC   ATCTGC-ACGTAC   ATCTG--ACGTAC
ATCTG-aACGTAC          ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG-CACGTAC   ATCTG-CACGTAC

Missing data : ? = A/T/G/C/-
Recoded:
ATCTGCCACGTAC
ATCTG--ACGTAC
ATCTG??ACGTAC
ATCTG??ACGTAC

Optimizations:
ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC
ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC
ATCTG--ACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGC-ACGTAC   ATCTGCCACGTAC
ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTGCCACGTAC   ATCTGC-ACGTAC   ATCTGC-ACGTAC

Polymorphic coding: 1 = C/-; 2 = A/-
Recoded:               Optimizations:
ATCTGCCACGTAC          ATCTGCCACGTAC   ATCTGCCACGTAC   ATCTGCCACGTAC
ATCTG--ACGTAC          ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC
ATCTG11ACGTAC          ATCTGCCACGTAC   ATCTG--ACGTAC   ATCTGC-ACGTAC
ATCTG22ACGTAC          ATCTG--ACGTAC   ATCTG--ACGTAC   ATCTG--ACGTAC

Exclusion              Contraction : 1= C/-; 2= A/-
Recoded:               Recoded:         Optimizations:
ATCTGACGTAC            ATCTGCACGTAC     ATCTGCACGTAC    ATCTGCACGTAC
ATCTGACGTAC            ATCTG-ACGTAC     ATCTG-ACGTAC    ATCTG-ACGTAC
ATCTGACGTAC            ATCTG1ACGTAC     ATCTGCACGTAC    ATCTG-ACGTAC
ATCTGACGTAC            ATCTG2ACGTAC     ATCTG-ACGTAC    ATCTG-ACGTAC
```

Figure 3-2. The effect of character coding on homology statements in questionably aligned DNA sequences, using hypothetical sequences of four taxa. Underlining: violation of test of conjunction. *Italics*: character states not found in original data. Gray background: switching positions of the respective bases results in further possible optimizations with corresponding consequences for homology. For details see main body of text.

together using parsimony. Essentially, elision is a weighting procedure, giving characters with unequivocal position more weight than those for which the specific similarity statement is uncertain. Those characters with unambiguous alignment are found multiple times, whereas any particular type of a column in regions of ambiguous alignment is included with a lesser frequency (four times versus once in Figure 3-2). Character weighting is highly controversial, because it introduces subjectivity into the analysis (Wheeler, 1986). Elision also has effects on homology (Wheeler *et al.*, 1995: 5-6): “The elided data ... have the disturbing property of assigning multiple primary homologies to the same datum. ... the implications for homology are unsettling, since individual bases must have individual histories, but are not treated as such.” To rephrase their finding, the test of conjunction is failed for all positions with variable alignment. All other single positions, which occur more than once, are regarded as instances of character weighting and not failure of the test of conjunction.

Case sensitive

PAUP (Swofford, 1993) can be instructed to treat characters in a case-sensitive manner. Unequivocal primary homologies are shown in upper case, those with uncertain alignment in selected taxa are in lower case. Lower case states are then treated as missing data (coded ‘?’), but preserve the original information in a convenient form (Figure 3-2: Case sensitive, Recoded). The first problem with case-sensitive coding is the inability to express a gap in lower case, for which reason an *a priori* alignment must be chosen subjectively, unless gaps are specified as missing characters. The latter is clearly inappropriate, as pointed out above; in the present example it would result in a CC optimization for all taxa, i.e., render the character uninformative (not shown). Assuming gaps are treated as a fifth character state, lower case coding becomes pointless, as a particular position for the lower case characters must be chosen to begin with. Additionally, the lower case (= missing) characters can now take on any state (A, T, G, C, -). Character-state optimization will only consider existing states: C and -. The observed state A in taxon 4 will never be found in the optimization; only base C. If in taxon 3 or 4 a gap should be optimized, the test of conjunction is failed (Figure 3-2: Case sensitive, Optimizations). Case sensitive coding confuses the uncertain expression of a state with the uncertain position of an observed state in the sequence.

Missing data

If a region of ambiguous alignment is coded as missing data (= ?; *e.g.*, Whiting *et al.*, 1997), the parsimony algorithm is offered more possibilities to optimize the states for these missing data entries. Such a strategy seemingly offers flexibility but comes at a high price, because the same problems with character-state optimization as discussed above apply. One additional type of inconsistency can result in taxon 4, when CC is optimized; it fails the test of conjunction and contradicts the original observation at the same time (Figure 3-2: Missing data, Optimizations: italicized and underlined).

Polymorphic

Regions with ambiguous alignment can be coded as polymorphic, restricting the possible states a particular position may exhibit (see Wiens, 1995, for overview). Positions six and seven for taxon 3 are coded as 1 = C or -, and for taxon 4 as 2 = A or - (Figure 3-2: Polymorphic, Recoded). The possibility of assigning unobserved character states to a position is barred, but still only existing states (C or -) will be optimized. Taxon 4 will always show a double gap, hence, fails the test of conjunction consistently.

Taxon 3 will also fail the test of conjunction in two of the four possible optimizations (Figure 3-2: Polymorphic, Optimizations).

Exclusion

Questionably aligned regions may be excluded from analysis (Figure 3-2: Exclusion, Recoded. *E.g.*, Gatesy *et al.*, 1993; Cerchio & Tucker, 1998). In this example used, this method will result in the example used here in a sequence with only uninformative characters, leaving the relationship of the four taxa unresolved. The exclusion of characters leaves the homology concept intact. Specifically, some observations are disregarded because they are already accounted for as being not due to common ancestry. Therefore, these observations can be of no cladistic interest. The other strategies discussed above lead to the same unresolved topology using elaborate coding schemes, which as a consequence contradict the observation and/or fail the test of conjunction during character-state optimization. The only way to include observations whose similarity is highly doubtful is to classify them as entirely different entities. This translates to the introduction of additional character states. In its most extreme form, or from the perspective of a skeptic, such practice will lead to a data matrix composed only of autapomorphies.

Contraction

This less stringent method provides hypotheses of homology at a higher level of generality. The questionably aligned positions six and seven are combined into a single character. Taxa 1 and 2 are coded straightforwardly as C and -, respectively. For taxa 3 and 4, polymorphic coding is employed again (Figure 3-2: Contraction, Recoded). As mentioned above, character state optimization is restricted to existing states. Taxon 4 will always show a gap, taxon 3 either a C or -. The test of conjunction is failed in neither case, nor do the optimized character states contradict the original observations: homology is intact (Figure 3-2: Contraction, Optimizations). A conceptually similar approach has been proposed by Wägele (1994).

The reduction of two or more positions to a single one may raise questions regarding weighting issue. The problem of character weighting has only bearing if two otherwise equivalent coding schemes are compared. The latter is not the case here. Data contraction is carried out because of problems relating to representation of observational similarities. Hence, the character weighting argument, despite in general being of legitimate concern, has no force in the current context.

HOMOLOGY IS SPECIAL SIMILARITY *SENSU* REMANE (1952)

Cladistics is based on characters that share special similarities as opposed to overall similarities. Does every observation qualify as special similarity? Hardly so, as it is well-known from the classic insect wing - bird wing example. Nevertheless, Gatesy et al. (1993: 156) have used the total evidence argument to retain all available data at any cost: "... Kluge's (1989) notion of 'total evidence' should be extended to the use of scrambled alignment regions in phylogenetic reconstruction." The total evidence argument, however, refers to characters as used in cladistics, i.e., those harboring special similarity, hence, does not justify indiscriminant inclusion of observations. The inclusion of highly ambiguously aligned sequences should be avoided. The difficult question of where to draw the line between unequivocal and ambiguous alignment will be discussed in the following section.

Belief formation

Character 'selection' is a critical part in any cladistic study, because we select those states that require explanation by way of common ancestry. Observations of structures considered homologous are by definition explained first in elementary hypotheses of primary homology and subsequently in the composite hypothesis. Shared similarity must be evaluated at each level of generality and follows the division of whole and part. Although a skull (whole) is readily recognized as a homologous structure in all vertebrates, each bone of the skull (parts) has to be evaluated individually. Conversely, however, if all bones of the skull (parts) are considered to be homologous, then the skull (whole) must be homologous as well (see Mahner, 1998, on the intransitive nature of the part-whole relationship). Similarly, each base in a homologous gene must be critically assessed. The assessment comprises both the identification of the character state (red/green/blue, A/G/C/T/-), as well as their classification in a character (petal color, position 213), termed 'character-state identity' and 'topographical identity' by Brower & Schawaroch (1996), or 'propositional belief formation' and 'objectual belief formation' in epistemology (Audi, 1998). This characterization is fully compatible with the concept of the predicate language for observations discussed above. The character is the noun, the state its predicate. In morphology, the categorization of the individual conditions into discrete states is more problematic than with molecular features, but the identification of the character is a greater problem with molecular characters. We readily classify the observed appendage of a tetrapod as a leg (topographical identity), but struggle with the description of it as stout or slender (character state identity). With molecular data, the identity of the character states (A, G, C, T, -) are unequivocal, but the position (topographical identity) on the sequence is to be determined. Character state identity and topographical identity must be satisfactorily assessed in order for an observation to be considered worthy of an explanation as shared similarity. The equivalent treatment of all characters is maintained, although different problems regarding the establishment as shared similarities are encountered (Brower & Schawaroch, 1996).

In cladistics, the characters of interest to us are restricted to those which we consider worthy of the explanatory effort. The cut-off point, whether or not one may still suspect certain structures as similar and to be explained in terms of primary homology in the elementary hypothesis, depends on the trust in one's own observations (Ax, 1989). Remane (1952:103) phrased it clearly: "Because of this methodology it follows that ... a transitional area exists, in which a carried out or doubted homologization depends on the optimistic or pessimistic temperament of the researcher" [translation from German]. This position is not unique to phylogenetics, but a general, philosophical/psychological principle. It is in the nature of belief formation that "[p]eople differ markedly in the beliefs they form about the very same things they each clearly see" (Audi, 1998:17). The boundary between recognized and doubted shared similarity is not sharp and no hard and fast rules exist as how much uncertainty is sufficient during the process of belief formation to argue for the *a priori* non-inclusion of a morphological character (Remane, 1952) or the *a posteriori* exclusion of a molecular character (see also Gatesy *et al.*, 1993) relative to the time of data acquisition. The treatment of the *a priori* exclusion of morphological characters and the *a posteriori* exclusion of molecular characters is equivalent; the perceived difference is inherent to the mode of data acquisition and is unrelated to the establishment of primary homologies.

The entire process of belief formation using auxiliary information is but another form of inference to the best explanation, i.e., abductive (Devitt, 1997). Abduction can not furnish an explanation in the form of newly postulated past ancestors for the present observations that is certain to be true, even if all the premises are true; as a form of hypothetic reasoning, abduction is ampliative and not truth-preserving. This is in marked contrast to valid deductive reasoning, where any conclusion must be true if the premises are true (non-ampliative, truth-preserving). The very nature of explanatory hypotheses is the inclusion of non-observed causal entities, in this case common ancestors, to account for observed effects, shared similarity. Any conclusion in the form of an hypothesis from an abductive inference, elementary as well as composite, will always be tentative.

Rational for data exclusion

To exclude characters after their acquisition was characterized by Wheeler (1986:108) as “to give up” with a certain, highly homoplastic data set if even reluctantly applied weighting did not provide better resolution. In the same sense, if multiple weighting schemes in sequence alignment lead to different similarity statements, then we may well give up these characters, i.e., exclude them. From an explanatory point of view, these observations are judged not worthy of explanation. Although unsatisfactory, it is more honest to admit the failure to recognize shared similarity, than to make unfounded assertions. A primary homology statement indicates that two or more properties are considered by an investigator sufficiently similar in terms of structure and position as to be tentatively accounted for by common ancestry. As a result, responsibility and accountability is bestowed back upon each practicing systematist. It is not the computer that proposed a particular hypothesis of relationship—and many ‘intriguing’ phylogenies have been published particularly with molecular data—, but informed investigators stand behind them. Some of the more eggregious examples from molecular studies include the following: non-monophyly of Tracheata (Ballard *et al.,* 1992; see Wägele & Stanjek, 1995; Farris, 1998); sperm whale as sister taxon to baleen whales:

(Milinkovich, 1995; see Heyning, 1997; Messenger & McGuire, 1998; Cerchio & Tucker, 1998); position of Pogonophora (McHugh, 1997; see Siddall *et al.*, 1999).

It has been argued that the exclusion of molecular characters is a cardinal sin in cladistics (Gatesy *et al.*, 1993). What is the comparable procedure with morphological characters? It would be the non-inclusion of available characters. As an example, in a class-level analysis body color of the exemplar taxa is not found in the data matrix (*e.g.*, Whiting *et al.*, 1997), although this character is readily observed. Why is this character not included? The inexplicit answer is that the primary homology of the character 'color' is questionable at the class level, although it is useful at the genus/species level (*e.g.*, Westerneat, 1993; Swenson & Bremer, 1997). From an observational point of view, the observer considers the superficial similarity unconvincing, so that no case for special similarity can be made. From an explanatory point of view at the level of the elementary hypothesis subsequent to the observational phase, the character-state distribution is already explained as being due to some causal event(s) other than common ancestry. The decision of inclusion or exclusion of a character is based on the investigator's willingness to account for shared similarity by way of common ancestry, by way of some event other than common ancestry, by way of not trusting their own observations, being unsure of what they observe, or by way of their inability to characterize what they observe. With morphological characters the arguments for non-inclusion are hardly ever spelled out, unless a previously used perceived similarity can no longer be accounted for by common ancestry at the level of the elementary hypothesis, i.e., is no longer considered a primary homology.

Similar practices can be found with sequence data. The use of the ITS region for population studies and species-level investigations is common (*e.g.*, Vanherwerden *et al.*, 1998; Mes *et al.*, 1997; Nakasone & Sytsma, 1993). Expanding the scope of such a study to phyla will result in unalignable sequences of approximately 25% similarity, which is tantamount to the random similarity of sets composed of four elements (Li & Graur, 1991). It is more sensible to select a more appropriate gene/gene-region that will allow for better alignment, than to argue for inclusion of all available data. Hence, we should try to avoid the problem that Wenzel (1997:37) characterized pointedly as "Garbage in, garbage out." The example chosen is extreme, but illustrates the point to be made at any level of taxonomic inclusiveness. Here then another perceived, major advantage of sequence data is put into perspective: the large number of characters obtained from sequences. As only synapomorphies are of explanatory interest, variable regions furnish such characters. However, it is also variable regions where the problematic alignments, i.e., designation of shared similarity, are prevalent, therefore, where the determination of primary homology is much more uncertain. The exclusion of questionably aligned regions also reduces the number of explanatorily relevant, what most refer to as 'informative', characters significantly, which may also lead to a loss of resolution in the cladogram. The potential to increase resolution has led to the use of methodologies in clear conflict with the philosophical basis of homology; Wheeler *et al.* (1995: 3) noted that "... the elided result was much more resolved", but also came to the conclusion that, "[c]learly, these are not the best of data to resolve insect relationship." The resistance to exclude characters can be followed, but does not provide a justification to include characters of doubtful primary homology, because the goal of cladistics is explanation of character state distributions, and not maximum resolution of its representation, tree topology. Note, that there is no clear-cut argument for the inclusion or exclusion of any particular character, because this decision relates to the process of belief formation and is inherently subjective (see previous section). There will never be a panacea for this problem, but being aware of this difficulty may help to avoid some major pitfalls. It is important to realize that an uncertain explanation of character-state occurrence shown as an unresolved part of a tree is increasing our knowledge about the cause of the observed distribution. By not knowing something causally, we automatically do know something: we recognize our ignorance, hence, we have further direction for our research (cf. Wenzel, 1997).

With molecular data, the uncertainty of topographical identity can be quantified. Positions that are not affected by any set of alignment parameters are identified as unequivocal similarities. Those positions that are stable over a wide range of alignment parameters are less equivocal than those, which change positions with any alignment parameters. Hence, a minimum range (*e.g.*, gap cost ratios from two to ten) over which the alignment must be unambiguous can be defined. Those observations remaining unambiguous are accepted as shared similarities, the remainder are excluded. The conditions (gap-cost ratio two to ten) are explicit, but still subjective. As long as the evaluated alignment parameters do not cover the entire spectrum of possible values (0.5 to infinity) the alignment is based on assumptions which are inherently subjective. Whether it is more appropriate to use an explicit, but rigid exclusion argument, than to exclude characters with implicit, but flexible exclusion arguments is open to debate. After all, what we are interested are observations in need of causal explanation by means of common ancestry.

One may argue that as the number of taxa in an analysis increases, the portion of the sequence with questionable alignment increases. Hence, all characters need to be included as eventually any position will be questionably aligned. The above is an oversimplification of two distinct cases. When taxonomic sampling density is increased, more landmarks will be recognized, which will aid in positioning the variable regions, because the additional taxa provide more auxiliary information to be used in the process of belief formation. These additional data furnish new arguments for the alignment of a given set of sequences; both resolution of previously questionable alignments as well as flagging of ‘new’ uncertain primary homologies are possible (Eernisse, 1997). With increased taxonomic sampling width, however, the new auxiliary information will not provide the landmarks needed to position variable regions. The tentatively identified similarities can not be upheld at that particular level, which is related to the explanatory relevance of information.

HIGHLY DISSIMILAR TAXA

What is the most appropriate action if only one or a few taxa are extremely dissimilar as compared to the remainder (Figure 3-3)? Often ‘dissimilar’ is equated with ‘divergent’, although the latter implies an explanation at the outset. Exclusion or contraction of data may eliminate much important information on the relationships of the majority of taxa and would reduce the data matrix to the lowest common denominator. One may consider missing-data coding for that particular stretch in highly dissimilar taxa (Figure 3-3: taxa 6 and 7), because essentially one does not know anything about the specific homologies in these taxa. However, as pointed out above, such a coding strategy is inappropriate because it confuses the uncertain expression of a state (= character state identity) with the uncertain position (= topographical identity) of an observed state in the sequence and will create problems during character-state optimization. The two coding strategies introduced below address the question of how to best represent our observations. As belief formation is a psychological problem beyond scientific mechanics, there is no conclusive answer.

Original data	Stretch coding		Block coding			presence/absence	
ATGTGCCACGAAC	ATGTGCC	ACGAAC	11	ATGTGCC	ACGAAC	ATGTGCCACGAAC	000000
ATGTGCCACGAAC	ATGTGCC	ACGAAC	22	ATGTGCC	ACGAAC	ATGTGCCACGAAC	000000
ATCTG--ACGAAC	ATCTG--	ACGAAC	33	ATCTG--	ACGAAC	ATCTG??ACGAAC	000011
ATCTG--ACGAAC	ATCTG--	ACGAAC	44	ATCTG--	ACGAAC	ATCTG??ACGAAC	000011
ATCTGAAACGTAC	ATCTGAA	ACGTAC	55	ATCTGAA	ACGTAC	ATCTGAAACGTAC	000000
-T----AACGTAC	6666666	ACGTAC	TA	6666666	ACGTAC	?T????AACGTAC	111110
-T----AACGAAC	6666666	ACGAAC	TA	7777777	ACGAAC	?T????AACGAAC	111110

Figure 3-3. Coding strategies for a few highly dissimilar taxa 1-7 shown in rows 1-7. Original data shows one of the many possible alignments of this particular data set. Stretch coding and block coding illustrate two alternative coding strategies compared to presence/absence coding discussed in the main body of the text.

Stretch coding

In Figure 3-3, the observed dissimilarity is a piece of available information about taxa 6 and 7 to be represented as such. This information can be expressed using new character states, i.e., the entire unalignable region can be coded as a single state 6 (Figure 3-3: stretch coding). In this fashion it can be shown that taxa 6 and 7 are both highly dissimilar and show synapomorphies for the first seven characters. Some may argue that the statement of high dissimilarity should just be coded in a single character, essentially contracting the seven positions. Others may say that major differences are seen in all seven characters for which reason stretch coding shows the condition appropriately. It may seem as if a weighting issues arises. It has been shown above (section Contraction) that the representation of observational similarities in agreement with homology takes precedence over considerations of weighting.

Minor differences between taxa 6 and 7 could be coded using an additional character state for taxon 7 (not shown). One caveat applies: as additional state(s) must be specified for a particular character, it implies a homology statement with the characters in taxa 1 to 5, which is not feasible. Explicit homology statements of T and A in taxa 6 and 7 are made, disregarding alternative alignment possibilities, *e.g.*, the T may either be found in position 2 as indicated or in position 4; the relative position of T may not be the same for taxa 6 and 7. One may argue that *shared* character *states* need to be explained and that the classification of the character states under a particular character is of lesser importance, because the homology statements are restricted to within the blocks of taxa 1 to 5 and 6 and 7 (Figure 3-3: boxes). Then the information is sufficiently represented by stretch coding.

Block coding

To circumvent the problem of unspecifiable homology statements across blocks within a stretch (Figure 3-3: Original data: characters 1-7, taxa 1-5 and characters 1-7, taxa 6-7), one may consider treating states in taxa 1-5 and 6-7 separately (Figure 3-3: block coding: right boxes), inserting autapomorphies for the corresponding taxa in the other block (Figure 3-3: block coding: left boxes). There are two blocks with autapomorphies only (characters 1-2, taxa 1-5: upper left box; characters 3-9, taxa 6-7: lower right box), and two blocks containing the sequence information (characters 3-9, taxa 1-5: upper right box; characters 1-2, taxa 6-7: lower left box). The homology statements of the sequence information in each block are unconnected to one another. No homology statements are made between the T in taxa 6-7 with respect to any T (in position 2 or 4 of original data) in taxa 1-5. However, problems with character weighting arise because the information from the first 7 characters is now coded in 9. The above discussed issue of homology statements within the lower block also apply here, i.e., are the T's and the A's in taxa 6-7 homologous or not? One may argue that the autapomorphies in respective blocks should be coded the same within each block but different from the information-bearing entries. Such coding would accentuate the between-block differences beyond the original observations, which is not supported here.

Either stretch- or block-coding strategies have theoretical advantages and disadvantages, resulting in a classical trade-off situation inherent to the psychological process of belief formation. Block coding may seem somewhat preferable for the following reasons. When considering taxa 1-5 and 6-7 separately, then the homology statements are clear. Only by combining the data might problems arise. Therefore, the information from the two blocks should be coded separately. As the other block does not contribute any information to the data found in the block under consideration, the ‘empty’ blocks should be assigned uninformative autapomorphies.

Presence/absence coding

Block coding is somewhat similar to the established procedure to code gaps in a supplemental presence/absence (p/a) matrix (e. g., Baum *et al.* 1998) (Figure 3-3: presence/absence). The two differ twofold. 1) The characters with gaps in the p/a strategy are left in the data matrix, whereas block coding recodes the characters within the sequence portion. Issues with character weighting are reduced to a minimum with block coding. 2) The gaps in the sequence part of the p/a matrix are treated as missing characters, whereas no missing character states are introduced with block coding. As any missing character state misrepresents an actual observation as the result of belief formation, any use of ‘?’ as an indicator of uncertain assignment of observed character state (= character-state identity) to a specific character (= topographical identity) is positively misleading. P/a coding is also at odds with the classification process in the observational phase. By establishing separate bins (characters) for observations actually belonging in one and the same bin, arguments for the actual existence of separate classifications are introduced: parallel explanatory universes are in effect advocated and this results in a schizophrenic view of the real world. Devitt (1997) and Mahner & Bunge (1997) discussed in much detail the need of a scientist to be a realist.

Comparison to practice in morphology

Comparison of the coding strategies (particularly stretch coding) to practices in morphology is favorable. In morphology, a particular state recognized to be classified in a particular character that is highly dissimilar in a particular taxon is coded as a separate state and is not forced into an existing state. One practical difference between morphological and molecular characters must be discussed.

A single observation can be coded as an additional character state in morphology but not with molecular data. With sequence data the problematic part in the process of belief formation is the establishment of topographical identity, therefore, a single position can not harbor ambiguous alignment. The minimum number of characters required for ambiguous alignment is two adjacent positions. If the auxiliary, positional information places a base in a particular position (= topographical identity), the identification of the character state (= character-state identity) is not an issue. For morphological characters, in contrast, the problematic part in the process of belief formation is character state identity. Hence, for any single set of observations classified in a character, any particular observation may not be classifiable in one of the other states. An observation that can not be classified in an existing state is given a new state. The common denominator is, that whenever one of the two conditions needed to postulate shared similarity is not met, topographical identity or character-state identity, then additional character states are introduced. In neither case are the particular characters coded as missing data, because otherwise the original observation can be contradicted during character-state optimization.

The distinction between topographical identity and character-state identity has further bearings. In morphology, a property of taxon X that can not be identified as a shared similarity, is given a new state, but remains classified in its original character. The new state introduced and found only in a single taxon does not indicate a relationship with the other taxa, except for ‘X not sharing a most recent common ancestor with any other taxon’ at the level of the elementary hypothesis. In essence, a questionable property is given a new state.

With molecular data, an identified state that can not be classified is assigned to a new grouping (position). The new character introduced shall only carry information for the taxon or taxa in which it occur(s). For the remainder of taxa, it should remain uninformative with respect to the expressed relationship at the level of the elementary hypothesis. This is best achieved using autapomorphies for all except the highly dissimilar taxon/taxa. For molecular data, then, a questionable position is given a new position. Hence, the introduction of a new position in block coding is related to the problematic topographical identity of readily recognizable states and is not a fundamental difference as compared to practices in morphology.

BETTER ALIGNMENT?

One could ask whether the above alignment strategies produce "improved representation of homology". The catch phrase in itself lets the old molecular jargon of '87% homologous' resurface, overlooking the binary nature of homology: is homologous, is not homologous. Hillis (1994: 339-340) was correct in stating that "molecular biologists may have done more to confound the meaning of the term homology than have any other group of scientists. .... Why this confusion of terms [homology versus similarity] has arisen in molecular biology is not clear; perhaps the term homology is thought to make the work sound more like science ...". Today a probabilistic notion has been added to the homology concept, particularly under maximum likelihood, an issue to be discussed elsewhere.

The question arises of how to evaluate the homologies. A number of avenues may be considered, which will be addressed below.

Comparisons to ‘known’ phylogenies: If a phylogeny would be known, then why bother trying to build a hypothesis using an ampliative, non-truth preserving mode of inference? One could only conclude that the mode of inference is properly characterized. It does not address the perceptional question in any form.

Conferred bootstrap, jackknife or similar support: It has been widely realized that these measures are fraught with problems of which Wenzel (1997) provided an overview. To justify one alignment over an other by means of a questionable metric is untenable. Determining nodal support indices for an entire tree has not even been attempted. All these metrics are based on the entire data matrix or permutations thereof, whereas justification of the alignment applies for every single position individually, respecting character independence and relevance of auxiliary information. Accordingly, any approach using support indices of any sort is misguided. If one would consider to use a character support index such as the rescaled consistency index, then the question arises, for which position the comparison is made, because the topographical identity of a perception is at stake.

Value of alignment score: To evaluate one alignment using other alignment parameters or other methods of alignment also misses the point. One would only compare the underlying models as characterized by the alignment parameters and could conclude that, indeed, they differ. The same problems with overall metrics and character-specific metrics arise as discussed in the previous paragraph.

As shown above, justification for a given alignment procedure can not come from the inferred hypothesis or any metric associated with the alignment. The justification needs to come from the factors surrounding the goal to be achieved, namely finding similarities worthy of explanation. It is accomplished by elimination of conflicts with the cladistic methodology at large: character independence and relevance of auxiliary information, as well as contradiction with observations and violation of the test of conjunction during character state reconstruction. These are the avenues that have been pursued, and on which the arguments have been built. Any criticism needs to address these issues.

CONCLUSIONS AND RECOMMENDATIONS

The following points can be extracted to form a guideline to DNA sequence alignment and character coding.

- Global alignment is inappropriate, because it conflicts with character independence. After conserved regions are identified, only characters between two adjacent, conserved regions should used to establish primary homologies of bases within: minimal fragment alignment (MFA) should be practiced.
- Objective sequence alignment is inherently impossible; some level of subjectivity is always introduced. Due to MFA, computer-facilitated and manual alignment each have their discrete advantages. The former allows for explicit specification of the assumptions in the form of alignment parameters used, and the latter forces one to critically justify every similarity statement in the data matrix.
- Gaps should be coded as a fifth character state as they are invoked during the process of belief formation in the observational phase.
- Flexible coding strategies (elision, case sensitive, missing data, polymorphic, presence/absence) all conflict with the test of conjunction or have the potential to contradict our original observations, hence, can not be justified in any explanatory context. Only data exclusion and data contraction do not introduce such problems.

- Data exclusion of highly ambiguously aligned regions is not in disagreement with the total evidence argument, as the latter applies only to special similarities *sensu* Remane (1952).
- For highly dissimilar taxa, new characters need to be introduced to represent the available information most appropriately. Highly dissimilar regions should be recoded with block or stretch coding.

Treating DNA sequences in such a fashion is fully compatible with coding strategies used for morphological data and assures compliance with fundamental principles of cladistic analysis, particularly the concepts of causal explanation and homology.