The "street light syndrome", or how protein taxonomy can bias experimental manipulations

Gabriel Markov,[1,2] Guillaume Lecointre,[3] Barbara Demeneix,[1] and Vincent Laudet[2]*

Summary

In the genomics era, bioinformatic analysis, especially in non-model species, facilitates the identification and naming of numerous new proteins, the function of which is then inferred through homology searches. Here, we question certain aspects of these approaches. What are the criteria that permit such a determination? What are their limits? Naming is classifying. We review the different criteria that are used to name a protein and discuss their constraints. We observe that the name given to a protein often introduces a bias for further functional analyses, a bias that is not often taken into account when analysing results. Last but not least, the heterogeneity of criteria used for naming proteins leads to self-inconsistent or contradictory protein classification that is potentially misleading. Finally, we recommend a wider use of phylogenetic criteria in protein naming.
***BioEssays* 30:349–357, 2008. © 2008 Wiley Periodicals, Inc.**

[1]USM 501, Evolution des Régulations Endocriniennes. Muséum National d'Histoire Naturelle, Paris, France.
[2]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, Molecular Zoology team, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland Lyon Sud, France.
[3]UMR 7138 CNRS-UPMC-IRD-MNHN-ENS CP26 Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Paris, France.
Funding agencies: We are grateful to Ecole Normale Supérieure de Lyon, Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique and the Ministère de l'Education Nationale, de la Recherche et de la Technologie for financial support. V.L. and B.D. laboratories are supported by the Cascade EU Network of Excellence.
*Correspondence to: Vincent Laudet, UMR 5242 du CNRS, Institut de Génomique Fonctionnelle de Lyon, Equipe de Zoologie Moléculaire. Université de Lyon, INRA IFR 128 BioSciences Lyon-Gerland, Ecole Normale Supérieure de Lyon 46, allée d'Italie 69364 Lyon Cedex 07, France. E-mail: vincent.laudet@ens-lyon.fr
DOI 10.1002/bies.20730
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: AR, androgen receptor; CRABP, cellular retinoic acid binding protein; DHA, docosahexaenoic acid; ER, estrogen receptor; ERR, estrogen related receptor; HSD17B, 17-β hydroxysteroid dehydrogenase; GR, glucocorticoid receptor; IkB, inhibitor of kappa B; MR, mineralocorticoid receptor; PR, progesterone receptor; RA, retinoic acid; RXR, retinoic X receptor; SDR, short-chain dehydrogenase/reductase; TPO, thyroperoxidase; USP, ultraspiracle.

Introduction

Among the many steps in describing a new protein, one that could be considered as trivial is its naming. The importance of this step seems to have been underestimated, as many examples show that giving a name to a protein is not neutral. Names refer to concepts and, therefore, could have a major influence on further experimental efforts.

In the pre-genomics era, when genes and proteins were isolated in order to understand the molecular basis of a phenotypic feature, the name chosen was often linked with the approach used to isolate the protein (a point discussed later). DNA probes were designed on the basis of known sequences, and used to search for new sequences hypothesised to be related enough to hybridize with the probe.

In the genomics era, the problem took on a new dimension with the increasing number of available nucleotidic sequences and the progress in prediction algorithms, which resulted in more and more new proteins, especially in non-model species, being predicted every day through bioinformatic analysis, with their function duly inferred by homology searches. Protein annotation is a two-step process (reviewed in Refs 1,2). First, there is a structural annotation step, in which the corresponding DNA sequence is checked for the presence of start and stop codons, splicing sites and other features that permit determination of the coding sequence. This step is now quite well automated (reviewed in Ref. 3), even if all prediction programs need to be refined, especially when working with sequences from non-model species, where the gene structure and splicing mechanisms can vary. The second step is the functional annotation of predicted protein products. This process is increasingly carried by automatic tools, which are very helpful for a quick description of large datasets, but are prone to artefacts and, in particular, can lead to propagation of annotation errors. The process has been discussed and reviewed by Valencia,[4] who argued for a reliability score assignment to sequence annotation in order to facilitate a critical appraisal of the information.

Any language needs concepts, which can be defined as classes of objects. A class is a set of assembled objects sharing a special property.[5] Any name of general use is primarily attached to a class, not to the object itself. For instance biologists do not need to name individual molecules; therefore any protein name actually refers to a class of material entities. When a new protein is described, the name refers not only to the molecules that were in the test tube of the describer, but also to all the molecules sharing the same amino acid sequence that could be found in the body of the animal from which it was purified, and even in the body of other animals of the same species. Giving a name to an object is therefore dealing with classifications: it is assigning the particular object to a set containing other objects sharing common properties (i.e. a class or a concept). These properties are always arbitrary, but problems arise when different properties are used to create non-overlapping kinds of concepts dealing with the same objects. Risks and imprecision ensue when using words in the wrong conceptual framework. For example ''algae'' is an ecological concept. It covers all living things having photosynthetic activity in aquatic environments. Using this term in a phylogenetic classification is potentially misleading. Errors result when we wrongly identify the nature of the concepts that we are using. If one selects ''algae'' to compare their DNA sequences while expecting homogeneity of ''algal'' sequences, one would be surprised to find phaeophycean sequences (brown algae) more similar to ciliate sequences than to green algae sequences.[6] Green algae sequences are more similar to those of land plants. Following on from these ideas, the goal of the present paper is to highlight the heterogeneity of properties chosen to create the concepts used for protein naming. This heterogeneity constrains experimental possibilities and creates misunderstandings: names are chosen according to various criteria and they are later wrongly understood as names referring to structure and origin (i.e. names given using phylogeny), as they should in any comparative approach in biology. In this regard, current protein-naming approaches are not based on a self-consistent classification of proteins.

Here we will review the different criteria that are used to name a protein, in order to pinpoint the limits of each of them. Then we will study how these names can influence further experiments, and we will finally discuss how such bias might be corrected.

Different types of names, but a common definition problem

Proteins are given different kinds of names, i.e. their names refer to heterogeneous concepts, depending on the approach used to isolate them. These different types are summarized in Table 1.

Many proteins studied by traditional approaches were given functional names (for the different meanings of the word ''function'' in this paper, see Box 1): when a protein was purified in order to understand the molecular basis of a given biological function, the name often referred to this function. The reference to this function often supposes that a particular organ exists in the animal having this protein. In some cases, this leads to anomalous situations. Many recent papers still refer to prolactin in teleosts (see Ref. 7 for an example), even if this protein could of course not stimulate lactation in species that do not have a mammary gland. In fact, in teleosts, prolactin is involved in osmoregulation and this probably represents the ancestral function of this protein. An interesting example of trying to correct such inaccuracies of nomenclature is the case of the WNT proteins. The first *wnt* gene was identified as a proto-oncogene, activated in response to proviral insertion of a mouse mammary tumour virus, and was named *int-1*. In *Drosophila*, where its mutation led to a wingless phenotype, it was named ''*Wingless*''. Later it was recognized that *int-1* and ''*Wingless*'' were homologous and that they belong to a large family of related glycoproteins. Since a wingless phenotype is nonsense in mammals, in order to simplify the nomenclature,

Table 1. Summary of the different types of names found in protein nomenclature

Notion alluded to in the name	Example
Biological function	Prolactine
Localisation in an organ	TPO (ThyroPerOxidase)
Mutant phenotype	USP (UltraSPiracle)
Ligand binding ability	RXR (Retinoid X Receptor)
Presence of a conserved domain	HOXB1
Position in a gene cluster	HOXB1, HOXC4
Biochemical function	HSD17B (17-beta HydroxySteroid Dehydrogenase)

The cited examples are discussed in text.

Box 1. Different meanings of the word ''function''.

The use of the word ''function'' can be confusing in protein biology, because it refers to different things. For the purposes of clarity, in this study, we will distinguish three kinds of ''function''.

Biochemical activity refers to the kind of biochemical reaction made by the protein, e.g. dehydrogenation, for an enzyme, or to the type of modification undergone by the protein, e.g. ligand binding for a receptor.

Biochemical function is the reaction made in-vivo by the protein, e.g. dehydrogenation of a 17-beta steroid, binding of retinoic acid.

Biological function is the function in which the protein is involved at organism level, e.g. steroid biosynthesis, metamorphosis.

the whole family was renamed Wnt, a amalgam of Wingless and Int (reviewed in Ref. 8). In other words, the statement that homologous proteins were given names referring to two different frameworks (developmental for *Drosophila*, oncogenetic for mouse) lead to the formation of a framework-neutral name, suitable for both proteins.

When the differences between species are not so great, the given name sometimes refers to a supposed homology between two different species. Defining the peroxidase expressed in the endostyle of *Branchiostoma belcheri* as BbTPO (for *Branchiostoma belcheri* ThyroPerOxidase) suggests that amphioxus endostyle is homologous to vertebrate thyroid. For some authors,[9] the restriction of the expression of BbTTF-1 (thyroid transcription factor-1) and BbTPO to the endostyle strongly suggests that the endostyle is homologous to the follicles of the thyroid gland, whereas the traditional hypothesis is that the whole thyroid gland is homologous to the endostyle.[10] Thus, because BbTPO is considered a thyroperoxidase, its expression pattern provided evidence for the homology between two organs. In this case, the use of the same name is misleading, and even dangerous, because it favours a conclusion that has not yet been adequately tested. The name ''Endostyle Peroxidase'' would be more neutral in this case.

In the case of developmental genes, the name often refers to the associated mutant phenotype. For example, the name of the nuclear receptor USP was coined from the *ultraspiracle* phenotype, referring to the extra set of spiracles observed on larvae harbouring a loss-of-function mutation in this gene,[11] whereas the orthologous gene was named RXR for ''Retinoic X-Receptor'' in mammals, referring to its ability to bind retinoids.[12] In basal insects (excluding Diptera and Lepidoptera), the orthologous protein sequence was surprisingly more similar to vertebrate RXR than to *Drosophila* USP. Therefore, the homologous protein of these insects was given the name USP-RXR, even in the absence of any ultraspiracle mutant and even if the receptor was recently shown not to bind retinoid in vivo.[13] Purely, in terms of gene function, a third name would have been preferable, but the authors chose simplification and phylogeny as a guide for nomenclature.

Many protein names derive from the presence of a particular conserved domain. In vertebrates, the name of the famous HOX proteins refers to the existence of a conserved DNA-binding homeodomain, which in turn refers to homeosis (i.e. the transformation of one body part into another), observed when those genes are mutated. But this domain also exists in other proteins, like the gap gene products EMS or OTX in mammals, which are not called ''HOX'' because this name is reserved for proteins whose gene is located within a *hox* cluster.[14] In some cases, this definition is quite arbitrary: two *Evx* genes are located on the 5′-end of the mammalian *HoxA* and *HoxD* clusters, but *Evx* are not included in the *hox* genes category because its *Drosophila* ortholog, *Even-skipped* (*Eve*) is not located within a *hox* cluster.[14] Since the *hox* gene cluster in *Drosophila* appears quite derived when compared to those of other metazoans, this exclusion of *Evx* from the bona fide *hox* cluster may be considered as arbitrary.

The situation is much more complicated with protein names referring to biochemical functions. For example, the name 17-β hydroxysteroid dehydrogenase (HSD17B) is used for many proteins that are supposed to dehydrogenate 17-β steroids, an important step in steroid hormone biosynthesis. But, in fact, this name describes proteins with very different activities (reviewed in Ref. 15). Another problem is that all the proteins named as HSD17B are not members of the same protein family: the HSD17B5 is a member of the aldoketoreductase (AKR) protein family while the rest of the known HSD17B belong to the short-chain dehydrogenase/reductase (SDR) protein family. Even within the SDR family, the HSD17B-activity seems to have arisen several times independently[15] (see Fig. 1).

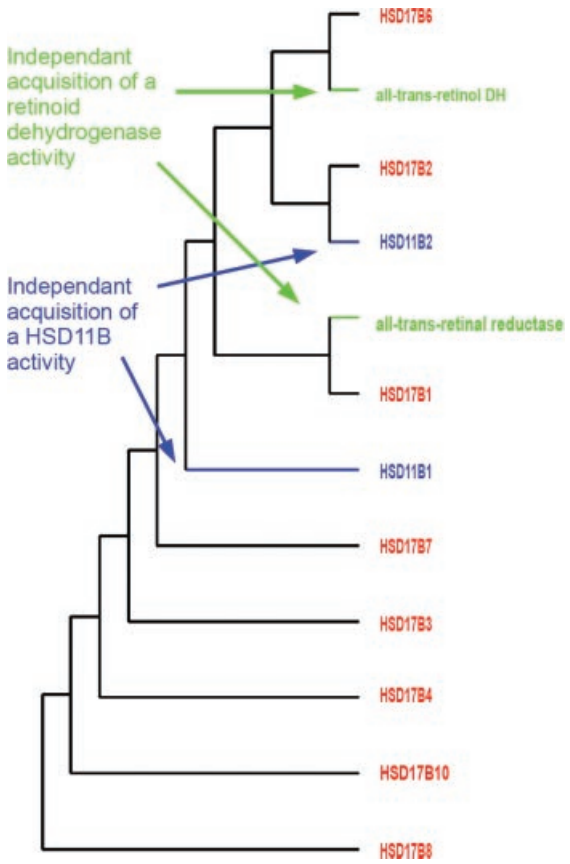

Figure 1. Convergent acquisition of biochemical functions within the SDR family in human. 11-beta-HSD activity and retinoid dehydrogenase activity seems to have appeared twice independantly. Because of the many other members of the family, not show, in that tree, even the HSD17B activity probably arose many times. Modified after Ref. 15.

These proteins often show multi-substrate activities, some of them being able to dehydrogenate either 17-β steroids or retinoids, with the HSD17B-activity sometimes only being established *in vitro* (reviewed in Ref. 16). The result is that the "HSD17B" family is paraphyletic, and even contains members with no 17-β hydroxysteroid dehydrogenase activity, leading to a very puzzling situation, since all these enzymes share the same identification code in the enzyme database (EC 1.1.1.51). This may also indicate that, in certain cases, the experimental efforts addressing the substrate specificity of the protein may have been misguided—or seen in a too narrow a fashion - by the gene name, as we discuss below. This provides a typical example of overlapping frameworks: the concept "HSD17B" was first used to describe a biochemical activity observed in vivo. But when the proteins presumed to be responsible for this activity were isolated, the word "HSD17B" was reused to name them, and was later used to name proteins showing an in vitro HSD17B-activity, whereas there was no evidence for their in vivo activity. So the same name "HSD17B" is used to describe two different concepts that partially overlap. The set "HSD17B" has a biochemical functional meaning. However, it is composed of entities that do not exhibit the biochemical functions in vivo because another framework, the presence of an in vitro HSD17B-activity, has led to their collation as a set. Such problems sometimes appear even in the title of the characterisation paper. For instance, the title "*Expression cloning and characterization of human 17 beta-hydroxysteroid dehydrogenase type 2, a microsomal enzyme possessing 20 alpha-hydroxysteroid dehydrogenase activity.*"[17] clearly indicates that assigning the name of a protein with referring to a biochemical activity lead to some overlaps, even at the biochemical level.

Furthermore, even when the biological function is conserved between two orthologous proteins, they can have radically different biochemical functions. The nuclear receptor USP-RXR is a transcription factor that is activated by the transient binding of small fatty acids in deuterostomes and molluscs. It has lost its ligand-binding pocket in insects, becoming an orphan receptor in most of the arthropods. But in Mecopterida, the crown insect group containing Diptera and Lepidoptera, the gain of a large ligand-binding pocket allows the binding of structural ligand.[18] In this case, fatty acids are constantly present in the ligand-binding pocket.[19] So even if the biological function of transcription factor, acting as a dimere with another nuclear receptor, is conserved among bilaterians, the biochemical function, here the ligand-binding abilities, are very different between Mecopterida, other insects and other bilaterians (Fig. 2). This difference cannot be overemphasised, because the ability to bind a ligand transiently allows fine-tuning of gene expression, and this will depend on the availability of ligand.

To sum up, protein nomenclature is very heterogeneous, often depending on historical circumstances and organism-linked specificities. For multigenic families, this often leads to confusing situations, so unified nomenclature systems have been established, specific for each gene family, to rationalize the system: *CYP* genes,[20] *hox* genes,[14] voltage-gated ion channels,[21] and nuclear receptors[22] provide classical examples. An independent nomenclature system has also been developed for all enzymes (reviewed in Ref. 23), official nomenclature committees such as NC-IUPHAR exist for receptors used in pharmacology,[24] and general nomenclature principles have been defined on a whole-genome scale for man.[25] These nomenclatures have their limitations, especially those based upon one specific organism (human or fly), because the different naming systems do not facilitate cross-species comparisons, and sometimes increase confusion. For example, Nelson recently warned about this situation giving the example of the annotation of rat *Cyp2* genes, where the nomenclature was fully revised by a nomenclature committee to match orthologous mouse genes, but the rat genes had already been given official names that refer to human *CYP* genes that are not their orthologs.[26] Another field that has received little attention is the correction of identified errors. It has been shown that a great number of papers are not retracted due to the lack of post-publication curation, especially for journals with low Impact Factor and limited access.[27] Thus, it would be important to facilitate post-publication correction in name fields in Genbank and other frequently visited databases. This implies that, when a protein name is seen to be erroneous, other curators should be able to submit modifications. As for accession number, former names should be conserved to permit an easy retrieval of sequences

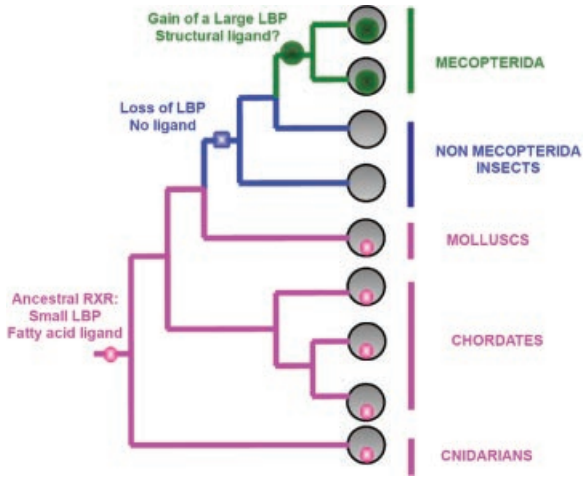

Figure 2. Function diversification in homologous proteins. The USP-RXR receptor binds small fatty-acid ligands in cnidarians, molluscs and chordates. It has lost its ligand-binding pocket (LBP) in insects and then gained a larger LBP in Mecopterida. Modified after Ref. 18.

referenced in old papers. Given the fact that such a work is time consuming, this curation activity should be taken into account both from a funding viewpoint and in the evaluation of researchers' activities. Securing an adequate funding for this often ignored, but critical, activity of curation will certainly have an important impact for the whole biology community since it could avoid controversies and experimental dead ends.

When protein names lead to experimental biases

One could argue that some names have only an historical signification and that their etymology has little importance for further studies. But selecting a name for a protein sometimes - and more often than expected—leads to experimental biases, simply because our experiments depend on the concepts we have in mind and sometimes these concepts are wrongly interpreted.

The estrogen-related receptor (ERR), was found using a low-stringency screen with a DNA probe corresponding to the gene region coding for the DNA-binding domain of human Estrogen Receptor alpha.[28] It was named ERR because of the high percent identity between its conserved domains (DNA- and ligand-binding domains) and the corresponding domains of estrogen receptor (ER). Given this proximity to ERs, its ability to bind estrogens was tested, and it was not possible to demonstrate binding with any major class of steroids.[28] It was later found that ERR-alpha can bind the endocrine disruptors toxaphene and chlordane[29] and that ERR beta and gamma are inhibited by 4-hydroxytamoxifen[30] even if the physiological relevance of these findings is still discussed.[31] But it has never been reported whether or not ERR could bind androgens, gluco- or mineralocorticoids, although new phylogenies suggest that ERR is not more closely related to ER than to other steroid receptors such as the glucocorticoid receptor (GR), mineralocorticoid receptor (MR), androgene receptor (AR) and progesterone receptor (PR).[32,33] The same bias appeared in studies on the biological functions of ERRs. Many features of ERRs (its DNA-binding site, its protein–protein interaction abilities, its implication in physiological pathways etc) were tested in the light of this apparent close relationship with ERs (see Ref. 34 for an example) and very little attention was paid to its possible links with steroid receptors even if we now know that ERRs are not more closely related to ERs than to other steroid receptors (Fig. 3). This bias clearly appears through a database search in PubMed abstract with the keywords ''estrogen-receptor related'' and other steroid receptor names. As of December 2006, only one abstract mentions together AR and ERR, two for PR and ERR, four for GR and ERR, none for MR and ERR, whereas 29 abstracts discussed relationships between ER and ERR, many of them with eloquent titles: ''*Transcriptional targets shared by estrogen receptor- related receptors (ERRs) and estrogen receptor (ER) alpha, but not by ERbeta*'',[34] ''*The mouse estrogen receptor-related orphan receptor alpha 1: molecular cloning and estrogen responsiveness*'',[35] ''*Estrogen receptor-related receptors in the killifish Fundulus heteroclitus: diversity, expression, and estrogen responsiveness*''.[36] This provides a striking example of a bias in the experimental effort, created by the use of a name derived from poorly resolved phylogeny.

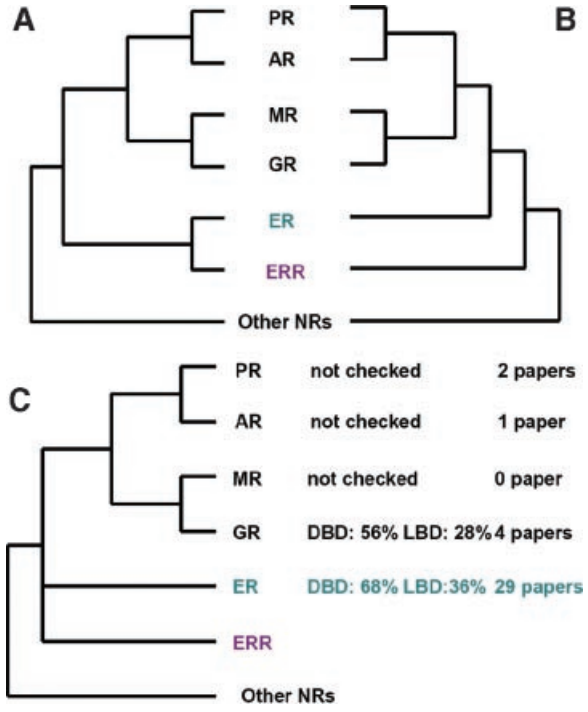

Figure 3. Protein naming and experimental effort. **A:** Phylogeny from Ref. 22, where ERR is closely related to ER, and thus explains why ERR was more often compared to ER than to other steroid receptors. **B:** Phylogeny from Ref. 32, showing that ER is not directly related to ERR, even if the last common ancestor of ER and other steroid receptors may have bound estrogens. **C:** Current consensus view of the phylogeny of ERR and other steroid receptors, showing that ERR is no more closely related to ER than to another SR. When provided by Giguere (Ref. 28), the percentages of amino-acid identity between the ligand-binding domain (LBD) and DNA-binding domain (DBD) of ERR and other receptors are shown. We also indicated the number of PubMed abstracts where comparisons are made between the ERR properties and those of other steroid receptors. There is a clear bias in favour of ER.

Such a bias also occurred in the initial studies on the nuclear receptor RXR. The name ''retinoid X receptor'' originally referred to its ability to bind vitamin A metabolites.[12] It was therefore supposed that RXR would be involved in a new retinoic acid (RA)-response pathway. Only recently, it was found that retinoids are apparently not *bona fide* natural ligand for RXR and that, in mouse brain, a fatty acid, the

docosahexaenoic acid (DHA) seems to be an endogenous RXR ligand.[37,38] But most of the papers studying RXR focus on its supposed involvement in RA-response pathway, whereas its involvement in fatty acid metabolism and signalling is much less studied, as shown in Table 2.

Another quite spectacular example is the case of the Cellular Retinoic Acid Binding Protein (CRABP) which are proteins implicated in retinoid signalling and related to FABP (Fatty Acid Binding Proteins). Retinoids are well known in vertebrates and their developmental role is well documented. But no defined retinoids have been isolated in arthropods and the existence of this signalling pathway in these organisms awaits clarification. The discovery of a protein that was supposed to be orthologous to vertebrate CRABP in the moth *Manduca sexta* was thus of interest. This first protein was cloned using a cDNA probe from a partial amino acid sequence of prothoracicotropic hormone that was similar to vertebrate retinoid-binding proteins. The newly isolated protein was annotated as CRABP in *Manduca sexta* on the basis of comparisons from percentage identities, without any real phylogenetic analysis.[39] The authors also proposed three-dimensional structures generated by homology-model building that showed the presence of an RA-binding pocket in the *Manduca* "CRABP". This protein was later used for the phylogenetic analysis of a newly cloned putative CRABP in the shrimp *Metapeneaeus ensis*.[40] The binding properties of the newly identified protein were also checked and it was concluded that the putative CRABP of *Metapenaeus* binds both RA and retinol, but not fatty acids (in fact the only fatty acid tested was parinaric acid). In 2005, a more exhaustive phylogenetic analysis of the family was performed, indicating that both sequences are members of the Fatty Acid Binding Protein (FABP) subfamily, a different subfamily, even if related to CRABP subfamily[41] (Fig. 4), showing that the genes coding for these insect proteins are not orthologous to the vertebrate CRABP. The binding abilities of *Manduca sexta* "CRABP" were also checked, showing that the *Manduca* "CRABP" has no significant affinity for RA and retinol, whereas it efficiently binds oleic acid and elaidic acid, and that the RA binding reported for *Metapenaeus ensis* seems due to experimental artefact.[41] This example shows that an excess of confidence in the result of an initial screen, together with a too-rapid phylogenetic analysis, can lead to a distorted experimental follow up. Moreover, the existence of the retinoid signalling pathway in arthropods remains elusive.

Ironically, within the same family, it should also be mentioned that a protein was annotated as CRABP in the amphioxus, *Branchiostoma floridae*[42] on the basis of a phylogeny using sequences from *Manduca* and *Metapenaeus* that initially annotated CRABP as an outgroup (Fig. 4). A more-detailed phylogenetic analysis (our unpublished data) suggests that this protein is probably neither a CRABP nor a CRBP but rather an IFAB (Intestinal Fatty Acid Binding Protein). This should be taken into account in further studies of its binding properties.

This example clearly shows that useful scientific papers can contain nomenclature errors. Such is the case from recent work on CRABP: the fact that the protein was not properly annotated makes not the crystal structure less interesting, even if the interpretation is different. It should be emphasized that the problem of protein naming is not only a problem of bad science, but that the use of functional naming is dangerous in itself. Other authors have already raised similar reservations and warnings on this subject (see for example Ref. 43 on bacterial *RecA*).

Table 2. Experimental effort about RXR

Key word searched	Number of abstracts registered in PubMed on January 3rd, 2007.
RXR Fatty acid(s)	197 (183)
RXR Glucocorticoids	25
RXR Retinoid(s)	1010 (1998)
RXR Steroids	362
RXR Thyroid hormones	176
RXR Vitamin D	315

The number of abstracts mentioning RXR with one of its putative ligand is reported. Even if recent experimental data seems to indicate that RXR is involved in fatty acid response, the number of abstracts mentioning RXR and Fatty acids together is quite low, compare to other metabolites, whereas the number of abstracts mentioning RXR and Retinoids is about ten times higher.

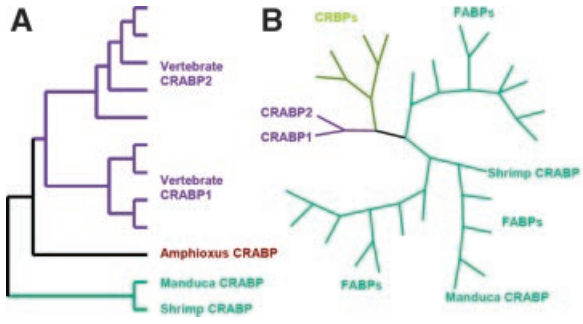

Figure 4. Influence of protein naming on phylogenetic sampling. **A:** Owing to the fact that some arthropod proteins were annotated as CRABP, Jackman et al. (Ref. 42) used them as an outgroup to check the phylogenetic position of a newly isolated amphioxus protein. **B:** Thus, Folli et al. (Ref. 41) showed that those proteins actually belong to the FABP subfamily, so these two proteins are not sufficient to indicate the position of the amphioxus sequence within the whole CRABP/CRBP/FABP family.

In conclusion, we need to emphasize that protein names with a functional significance are not neutral. They influence experimenters, and lead to what could be named the *street light syndrome:* as the joke goes, people tend to search for lost keys under the light of a lamp post, because it is more easy to search there than the actual place where the key was lost! Similarly, biochemists who study a known protein in a new species tend to check only if the protein has the functional properties that they suppose it should have on the basis of what is known about its function in other species, and not on the basis of protein sequence phylogenetic relationships.

One could wonder if such problems are also encountered in large-scale analyses, such as DNA arrays or EST analysis, where scientists have to rely on database annotations. Particularly in these cases, the name is generally not taken into account to infer protein functions. Orthology relationships are inferred from automated phylogeny pipelines, or filtered blast search results, and functional inferences are based upon Gene Ontology,[44] where indications about the reliability level of given information are indicated (for an example, see Ref. 45).

How can we avoid bias?

Even though functional names are insufficient to describe the real complexity of protein features, and can lead to experimental bias, manual correction of those errors represents a bottleneck, given the huge amount of new data. Moreover, powerful tools, such as Gene Ontology,[44] were developed specifically to provide functional information about a given protein, taking into account the reliability level of those data. So in future, it would be better to avoid functional naming of protein, in order to clearly distinguish between the name, which should be a constant tag used to refer to the studied object, and the description, which could evolve with the growing knowledge, and should provide the detailed features of the object. Such a distinction was already made between *signifiers* (i.e. names referring to an object) and *descriptors* (i.e. names referring to the known object properties), and some solutions have been proposed to avoid this confusion.[46] Basically, the proposal was to promote the use of gene names that do no refer to any describing features, either because they are only remotely connected to a mutation phenotype (for example *sonic hedgehog*), or because they come from languages other than English, and so have little signification for the majority of the research community (for example *fushi tarazu*). Generalising such a system could avoid many experimental biases but, however, will be of limited use for managing knowledge about big protein superfamilies. We want to emphasize that naming through a descriptor may not be problematic if the description method is universal and can be applied to all proteins. We propose that phylogeny, which is based on evolution, is an excellent tool to name genes with both accuracy and flexibility.

Indeed, as in systematics, object naming in molecular biology or in biochemistry has to manage structural knowledge. However, this is not enough. Molecular biology and biochemistry are sciences dealing with biological entities and phenomena; i.e. entities that vary through time and among populations and that have evolutionary histories. Therefore naming them according to phylogeny takes into account both structure and history, as clades are sets based on shared derived features. Biology has one general theory, evolution, and it seems appropriate and useful for all its sub-disciplines to take it into account.

Exhaustive phylogenetic analyses may also help to avoid many biases, and datasets should take into account not only the sister groups of the protein of interest, but also more distant families.

Too often, proteins are still annotated with BLAST,[47] which uses distance comparisons between sequences to make similarity scores. This implies that the newly identified protein is considered homologous to the most-similar protein available. The problem is that the first match is not necessarily the orthologous protein. The global similarity may not even indicate an orthology relationship. In multigenic families, one protein could be homologous to two different paralogous proteins in another species.[48] Fast-evolving proteins can also artefactually display great similarity simply because of random multiple substitutions at the same sites,[49,50] or because of similar composition biases.[51] Thus, only a careful phylogenetic analysis, taking into account these risks, can provide reliable information about the position of a new protein within a family. When trying to group different entities under the same name on the basis of their similarity, it is important to distinguish whether this similarity is a result of common ancestry of convergent evolution. Proteins are evolving entities, and undergo modifications of their features as a function of time, according to the diversification of the organisms to which they belong. Thus, an important concept is their evolutionary history, that is their relationships based on descent with modification and the inferred transformation events that they underwent.

Using phylogenetic taxonomy, the fact that some names do not fit the actual characteristics of all group members, due to functional shifts—at a biochemical or biological level—in different organisms, will not be too problematic (see Ref. 52 for an example how to detect these functional shifts). For example, the fact that snakes are tetrapods, even if they secondarily lost their legs, will probably not puzzle any zoologist, because ''tetrapod'' is an evolutionary concept. The word "tetrapod" is not used in a descriptive, fixist or essentialist meaning, but refers to the character states of the last common ancestor of this vertebrate group. Thus, observing that snakes have no limbs even if they are included within tetrapods

(because they do have other tetrapod features) gives the information that they secondarily lost their limbs. In the same manner, a phylogenetic classification of proteins may help to organise knowledge and to propose evolutionary hypotheses in the field.

Conclusion

We have reviewed many examples showing that protein names are not neutral and can lead to experimental biases. Common names are problematic because protein properties used in the past for creating concepts are heterogeneous, and therefore vary among species. Only a single conceptual framework for names can provide a self-consistent classification that the biochemists, geneticists and molecular biologists need. Further statistical studies could be useful to evaluate the global importance of this phenomenon, but it should be clear that protein names should no longer be based on heterogeneous concepts that narrow the experimental research field. As in systematics, concepts of monophyly should drive the attribution of names, because proteins are evolving entities.

To facilitate such applications we propose a number of suggestions, summarized in Box 2. Clearly, they are only starting points, not definitive methods, and a broad reflection and effort on this nomenclature problem are required to resolve it fully.

Note added in proof:

The ''streetlight syndrome'' should in fact be named the ''moonlight syndrome'', since it is already mentioned in a Middle East story about the mythic 13th century hero Nasreddin Hodja.

One night Nasreddin Hodja lost his ring down in the basement of his house, where it was very dark. Then he went out on the street and started looking for it there, under a splendid moonlight.

A friend passing by stopped and enquired:

- What are you looking for, Hodja? Have you lost something?

- Yes, I've lost my ring down in the basement.

- But Hodja, why don't you look for it down in the basement where you have lost it? asked the friend in surprise.

- Don't be silly, man! I prefer to search where there is some light!

Box 2. Preliminary suggestions to limit the nomenclature problems in protein taxonomy.

1. Official protein names should refer only to their phylogenetic relationships; only proteins from orthologous genes should have the same name; when it is not possible to make a detailed whole-family phylogenetic study, automated tools such as the curated database of phylogenetic trees of animal gene families, TreeFam[53] or the whole phylogeny pipeline http://www.Phylogene.fr should be used.
2. As for Gene Ontology, information about the reliability level of the name (manual curation, phylogeny pipeline, filtered BLAST search) should be mentioned.
3. Organism-based nomenclature should be abandoned.
4. In public databases, the naming field should be updated by database curators.
5. The naming effort should be supported by consequent funding.

Acknowledgments

We thank François Bonneton, Marc Robinson-Rechavi, the editor and three anonymous reviewers for their manuscript reading and useful critical comments, Frédéric Brunet and Michael Schubert for their fruitful discussions.

References

1. Rouze P, Pavy N, Rombauts S. 1999. Genome annotation: which tools do we have for it? Curr. Opin. Plant Biol 2:90–95.
2. Danchin EG, Levasseur A, Rascol VL, Gouret P, Pontarotti P. 2007. The use of evolutionary biology concepts for genome annotation. J Exp Zoolog B Mol Dev Evol 308:26–36.
3. Mathe C, Sagot MF, Schiex T, Rouze P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30:4103–4117.
4. Valencia A. 2005. Automatic annotation of protein function. Curr Opin Struct Biol 15:267–274.
5. Mahner M, Bunge M. 1997. Foundations of Biophilosophy. Berlin, Heidelberg, New York: Springer Verlag. p. 218.
6. Kumar S, Rzhetsky A. 1996. Evolutionary relationships of eukaryotic kingdoms. J Mol Evol 42:183–193.
7. Lee KM, Kaneko T, Aida K. 2006. Prolactin and prolactin receptor expressions in a marine teleost, pufferfish Takifugu rubripes. Gen Comp Endocrinol 146:318–328.
8. Siegfried E, Perrimon N. 1994. Drosophila wingless: a paradigm for the function and mechanism of Wnt signaling. Bioessays 16:395–404.
9. Ogasawara M. 2000. Overlapping expression of amphioxus homologs of the thyroid transcription factor-1 gene and thyroid peroxidase gene in the endostyle: insight into evolution of the thyroid gland. Dev Genes Evol 210:231–242.
10. Kluge B, Renault N, Rohr KB. 2005. Anatomical and molecular reinvestigation of lamprey endostyle development provides new insight into thyroid gland evolution. Dev Genes Evol 215:32–40.
11. Perrimon N, Engstrom L, Mahowald AP. 1985. Developmental genetics of the 2C-D region of the *Drosophila* X chromosome. Genetics 111:23–41.
12. Mangelsdorf DJ, Ong ES, Dyck JA, Evans RM. 1990. Nuclear receptor that identifies a novel retinoic acid response pathway. Nature 345:224–229.
13. Bonneton F, Zelus D, Iwema T, Robinson-Rechavi M, Laudet V. 2003. Rapid divergence of the ecdysone receptor in Diptera and Lepidoptera suggests coevolution between ECR and USP-RXR. Mol Biol Evol 20:541–553.
14. Scott MP. 1993. A rational nomenclature for vertebrate homeobox (HOX) genes. Nucleic Acids Res 21:1687–1688.
15. Baker ME. 2001. Evolution of 17b-hydroxysteroid dehydrogenases and their role in androgen. estrogen and retinoid action. Molecular and Cellular Endocrinology 171:211–215.

16. Peltoketo H, Luu-The V, Simard J, Adamski J. 1999. 17beta-hydroxysteroid dehydrogenase (HSD)/17-ketosteroid reductase (KSR) family; nomenclature and main characteristics of the 17HSD/KSR enzymes. J Mol Endocrinol 23:1–11.
17. Wu L, Einstein M, Geissler WM, Chan HK, Elliston KO, et al. 1993. Expression cloning and characterization of human 17 beta-hydroxysteroid dehydrogenase type 2, a microsomal enzyme possessing 20 alpha-hydroxysteroid dehydrogenase activity. J Biol Chem 268:12964–12969.
18. Iwema T, Billas IML, Beck Y, Bonneton F, Nierengarten H, et al. 2007. Ligand-Independent Functional Conformation of RXR-USP: Insight into Nuclear Receptor-Ligand Evolution. EMBO J 26:3770–3782.
19. Billas IM, Moulinier L, Rochel N, Moras D. 2001. Crystal structure of the ligand-binding domain of the ultraspiracle protein USP, the ortholog of retinoid X receptors in insects. J Biol Chem 276:7465–7474.
20. Nebert DW, Adesnik M, Coon MJ, Estabrook RW, Gonzalez FJ, et al. 1987. The P450 gene superfamily: recommended nomenclature. DNA 6: 1–11.
21. Ertel E, Campbell K, Harpold M, Hofmann F, Mori Y, et al. 2000. Nomenclature of voltage-gated calcium channels. Neuron 25:533–535.
22. Nuclear Receptors Nomenclature Comittee. 1999. A unified nomenclature system for the nuclear receptor superfamily. Cell 97:161–163.
23. Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res 28: 304–305.
24. Spedding M, Foord SM, Hofmann F. 2004. Current status of drug receptor nomenclature: receptor closure? The role of NC-IUPHAR. Expert Opin Investig Drugs 13:461–464.
25. Human Genome Nomenclature Comitee. 2002. Guidelines for Human Gene Nomenclature. Genomics 79:464–470.
26. Nelson DR. 2005. Gene nomenclature by default, or BLASTing to Babel. Hum Genomics 2:196–201.
27. Cokol M, Iossifov I, Rodriguez-Esteban R, Rzhetsky A. 2007. How many scientific papers should be retracted? EMBO Rep 8:422–423.
28. Giguère V, Yang N, Segui P, S Evans RM. 1988. Identification of a new class of steroid hormone receptors. Nature 331:91–94.
29. Yang C, Chen S. 1999. Two organochlorine pesticides, toxaphene and chlordane, are antagonists for estrogen-related receptor -1 orphan receptor. Cancer Res 59:4519–4524.
30. Tremblay GB, Bergeron D, Giguere V. 2001. 4-Hydroxytamoxifen is an isoform-specific inhibitor of orphan estrogen-receptor-related (ERR) nuclear receptors beta and gamma. Endocrinology 142:4572–4575.
31. Horard B, Vanacker JM. 2003. Estrogen receptor-related receptors: orphan receptors desperately seeking a ligand. J Mol Endocrinol 31: 349–357.
32. Thornton JW, Need E, Crews D. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301:1714–1717.
33. Bertrand S, Brunet FG, Escriva E, Parmentier G, Laudet V, et al. 2004. Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems. Mol Biol Evol 21: 1923–1937.
34. Vanacker JM, Pettersson K, Gustafsson JA, Laudet V. 1999. Transcriptional targets shared by estrogen receptor-related receptors (ERRs) and estrogen receptor (ER) alpha, but not by ERbeta. EMBO J 18:4270–4279.
35. Shigeta H, Zuo W, Yang N, DiAugustine R, Teng CT. 1997. The mouse estrogen receptor-related orphan receptor alpha 1: molecular cloning and estrogen responsiveness. J Mol Endocrinol 19:299–309.
36. Tarrant AM, Greytak SR, Callard GV, Hahn ME. 2006. Estrogen receptor-related receptors in the killifish Fundulus heteroclitus: diversity, expression, and estrogen responsiveness. J Mol Endocrinol 37:105–120.
37. de Urquiza AM, Liu S, Sjoberg M, Zetterstrom RH, Griffiths W, et al. 2000. Docosahexaenoic acid, a ligand for the retinoid X receptor in mouse brain. Science 290:2140–2144.
38. Lengqvist J, Mata DeUrquizaA, Bergman AC, Willson TM, Sjovall J, et al. 2004. Polyunsaturated fatty acids including docosahexaenoic and arachidonic acid bind to the retinoid X receptor alpha ligand-binding domain. Mol Cell Proteomics 3:692–703.
39. Mansfield SG, Cammer S, Alexander SC, Muehleisen DP, Gray RS, et al. 1998. Molecular cloning and characterization of an invertebrate cellular retinoic acid binding protein. Proc Natl Acad Sci USA 95:6825–6830.
40. Gu PL, Gunawardene YI, Chow BC, He JG, Chan SM. 2002. Characterization of a novel cellular retinoic acid/retinol binding protein from shrimp: expression of the recombinant protein for immunohistochemical detection and binding assay. Gene 288:77–84.
41. Folli C, Ramazzina I, Percudani R, Berni R. 2005. Ligand-binding specificity of an invertebrate (Manduca sexta) putative cellular retinoic acid binding protein. Biochim Biophys Acta 1747:229–237.
42. Jackman WR, Mougey JM, Panopoulou GD, Kimmel CB. 2004. *crabp* and *maf* highlight the novelty of the amphioxus club-shaped gland. Acta Zoologica (Stockholm) 85:91–99.
43. Courcelle J, Ganesan AK, Hanawalt PC. 2001. Therefore, what are recombination proteins there for? Bioessays 23:463–470.
44. Lan N, Montelione G, Gerstein M. 2003. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. Curr Opin Chem Biol 7:44–54.
45. Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.
46. Wilkins AS. 2001. Gene names: the approaching end of a century-long dilemma. Bioessays 23:377–378.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
48. Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99–113.
49. Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410.
50. Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7.
51. Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690.
52. Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, et al. 2006. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. BMC Evol Biol 6:92.
53. Li H, Coghlan A, Ruan J, Coin L, Heriche J, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:572–580.