Evolutionary Classification of Homeodomains

VIKTOR GINDILIS,^{1,2} EUGENE GOLTSMAN,¹ and YURY VERLINSKY¹

Submitted: January 7, 1998

Accepted: January 7, 1998

Purpose and Methods: We performed multiple comparisons between available amino acid (aa) sequences of homeodomain(HOM)-containing proteins from a wide spectrum of animals to create an evolutionary classification of the proteins.

Results: Based on results of statistical and special computational analyses of over 500 homeodomain aa sequences (HOMs) a novel system of concepts describing complex structural correlations between homologous proteins is proposed. This system includes such notions as differentiated isofunctionality of aa, chemotype, stereotype, local functional motifs, gradual conservativeness of aa positions, and group-specific domain patterns, as well as major categories of the evolutionary classification of HOMs (Division, Type, Branch, Class, Family, Series, Variety, Sort). Using this approach, a complete structural systematics of HOMs belonging to proteomes of eukaryotic animals is proposed.

Conclusions: The proposed structural classification of HOMs is in full agreement with the bulk of experimental data revealing complex functional similarities and differences among HOMs in terms of their expression patterns in developing embryos. It turns, this classification can provide answers regarding homology among homeodomains when experimental data are conflicting.

KEY WORDS: homeobox-containing genes; homeodomain-containing transcription factors; amino acid isofunctionality; homeodomain patterns; homeoprotein classification.

INTRODUCTION

Two major factors interfere with large-scale evolutionary studies of related genes and proteins, particularly of those controlling developmental processes. First, all public genetic sequence databases (GSDBs), both

nucleotide (nt) and amino acid (aa), contain a great amount of informational noise, i.e., multiple syntactical and semantic errors, as well as overredundancy of primary documents. Our earlier investigation of homeobox-/homeodomain (HOX-/HOM-)-containing nt/aa sequences submitted to the databases has revealed that 60% of GSDB primary documents (entries) cannot be directly used in evolutionary studies until the relevant information is refined from multiple syntactical errors and semantic deficiency/indeterminacy (1).

The second issue is a lack of a widely accepted system of categories that could be valid in defining groups of evolutionary related genes/proteins. For instance, terms such as gene/protein "class," "family," and "subfamily" are widely used in special literature, although strict definitions of the notions have not been presented, possibly by virtue of their "self-evidence" (2–4). Likewise, no one has discussed how many additional categories and which particular ones are needed to provide valid descriptions of complex structural and functional interrelations among many hundreds or even thousands of homologous genes/proteins.

Other important issues have also been raised recently in a letter published in *Trends in Genetics* by a group of developmental and evolutionary biologists concerned with conceptual difficulties that complicate correct interpretations of data obtained in comparative genomics of developmental mechanisms. The authors emphasized that "we must be clear at exactly what level we are inferring homology: genes, their expression patterns, their developmental roles, or the structures to which they give rise"(5).

Here we attempt to resolve some of these conceptual complications. Our innovations are based on a detailed study of a large group of developmentally important transcription factors—the HOM-containing proteins. A system of notions and terms describing complex structural correlations between homologous proteins is proposed, and special statistical and computational procedures are used to ground an evolutionary classification of HOMs.

¹ Reproductive Genetics Institute, 836 West Wellington, Chicago, Illinois 60657.

² To whom correspondence should be addressed.

MATERIALS AND METHODS

The nt and aa sequences of HOX/HOM-containing genes and proteins, as well as other relevant information, were obtained from original publications and from GSDBs—PIR, EMBL, SwissProt, and the NCBI's integrating database available under the *Entrez* online browser at the following Web sites: (<http://www-nbrf.georgetown.edu/pir/>), (<http://www.ebi.ac.uk/queries/queries.html>), (<http://expasy.hcuge.ch/sprot/sprot-top.html>), and (<http://www3.ncbi.nlm.nih.gov/Entrez/>).

To avoid the aforementioned informational noise, a program package was developed to refine and analyze primary nt/aa data (6). Overall, several thousands of primary documents submitted to GSDBs and/or obtained from original publications were studied. This allowed us to select approximately 500 HOM aa sequences.

In this report we restricted our discussion only to the HOM region, with its short flanking segments (in total, 69 aa; see Fig. 1), which resides, as a rule, in the carboxyl-terminal half of HOM-containing proteins. Since we showed earlier that the canonical HOM sequence consists of 57 aa (6), contrary to the commonly used 61–60 aa (2,3), Fig. 1 presents the three HOM scales in comparison. In addition, we modified an index “V” introduced previously in our studies to measure the variability of aa occurrences at each aa position within the protein domain (6). The modified index emphasizes the conservativeness of an aa position rather than its variability. Thus, if we had earlier $V = (1 - C)$, now we use the C index as follows:

$$C_i = \left\{ \sum_j^k (P_j - p)^2 / p \right\} / (k - 1)$$

Here C_i is the value of *conservativeness* for the i th aa position within a protein domain; P_j is the observed probability for the j th amino acid to occupy the i th position within a domain in a given sample of aa sequences; and p is the theoretical probability expected for any of the 20 common aa to occur at a particular position, in their random distribution within a protein domain. This C index ranges in value from 0.0, i.e., lack of conservativeness at a position, to 1.0, i.e., maximal conservativeness. As shown under Results and Discussion, we can increase the sensitivity of C by gathering aa into nine groups, based on their similarities. To distinguish the two indexes, we designated them sC_i for the case of *ungrouped* aa ($k = 20$) and gC_i for the case of *grouped* aa ($k = 9$).

It should be noted that a value of p used as the theoretical probability for any aa to occur at a particular position within a domain is derived from $1/k$; it will be 0.05 when $k = 20$ or 0.11 when $k = 9$. In actuality, this is an approximation because in different proteins the common aa occur at significantly different frequencies. We can accept the approximation merely for the purpose of a preliminary evaluation of the most conservative and the most variable aa positions within a protein domain. However, a better approximation could be used when one has the intention to use the C index in testing the statistical significance of data observed. We present and discuss this technique in Ref. 7.

RESULTS AND DISCUSSION

Key Concepts of Evolutionary Systematics of Protein Domains

First, we introduce a system of categories designed to provide an accurate description of the complex evolutionary correlations among HOMs with respect to their primary structures. The system (Homeodomain Evolutionary Systematics) uses three **basic** categories, *Division*, *Family*, and *Variety*—and several **optional** ones—*Type*, *Branch*, *Class*, *Section*, *Series*, *Sort*, and *Variant*.

Classification of HOMs into the categories is based on the concept of *domain pattern* (DP), especially group-specific DP, e.g., Division- or Family-specific ones. The purpose of these is to serve as identifiers of particular groups of homologous protein domains. To define the above concept clearly, we need to clarify some of the key terms: *chemotype*, *stereotype*, and *local functional motif*.

A *chemotype* is any relatively short continuous aa sequence (di-, tri-, tetrapeptides, etc.). For instance, the aa string “SELA” (Ser–Glu–Leu–Ala) that is present in several HOMs (Fig. 1; aa positions 31–34) is a chemotype; likewise, the aa string “WFQNR” (Trp–Phe–Gln–Arg–Arg; aa positions 47–52) is also a chemotype.

A *stereotype* is a 3D-configuration of a chemotype that is determined both by bonds between chemical radicals within a particular aa (i.e., *inner binding*) and by bonds between radicals belonging to different aa (i.e., *outer contacts*). For instance, each of the two mentioned chemotypes is involved in α -helix II or III, respectively, i.e., in specific stereotypes (see Fig. 1).

Ref.	2	Helix I										Loop			Helix II			Turn		Helix III/IV						GSDB ID
		6	9	11	13	17	21	23	27	29	31	36	38	41	43	51	59	61								
Ref.	3	5	8	10	12	16	20	22	26	28	30	35	37	40	42	50	58	60								
Ref.	6	-6	...	-1	+1	4	7	9	11	15	19	21	25	27	29	34	36	39	41	49	57	63				
dro.Cad	gktrtk	dK	yRvv	Y	tdf	Q	rle	L	eke	Y	ct	sr	YIt	ir	RR	se	LA	qt	LsLs	erqvki	WFQN	RR	aKe	Rt	snkks	X03062
mur.Cdx4	gktrtk	eK	yRvv	Y	tdh	Q	rle	L	eke	F	hc	nr	YIt	ir	RR	se	LA	vn	LgLs	erqvki	WFQN	RR	aKe	RR	mikkki	L08061
hum.CAB4D	ltaksg	RR	kRcp	Y	tkh	Q	tle	L	eke	F	lf	nm	Ylt	re	RR	le	IS	ks	VnLt	drqvki	WFQN	RR	mKl	KK	msrenr	X59373
mur.Cab5B	lharss	RR	kRcp	Y	tky	Q	tle	L	eke	F	lf	nm	Ylt	rd	RR	he	VA	rl	LnLs	erqvki	WFQN	RR	mKm	KK	mnngag	M34857
mur.AntA	ssgpd	RR	gRqt	Y	try	Q	tle	L	eke	F	hf	nr	Ylt	rr	RR	ie	IA	ha	LcLt	erqvki	WFQN	RR	mKw	KK	ehkdes	M17192
hum.PBP2B	pgsaas	RR	aRta	Y	tse	Q	lve	L	eke	F	hf	nr	Ylt	rp	RR	ve	MA	nl	LnLs	erqvki	WFQN	RR	mKy	KK	dqkag	X16667
mur.Mox1	egsska	RR	eRta	F	tke	Q	lre	L	eae	F	ah	hn	Ylt	rl	RR	ye	IA	vn	LdLs	erqvkv	WFQN	RR	mKw	RR	vkggqp	Z15103
dro.Eve	padpsv	RR	yRta	F	trd	Q	lgr	L	eke	F	yk	en	YVs	rp	RR	ce	LA	as	LnLp	estikv	WFQN	RR	mKd	RR	qriava	X05138
cen.Ceh13G5	gsseee	RR	pRta	F	tg	Q	ldr	L	kte	F	re	sr	Ylt	ek	RR	qe	LA	he	LgLn	esqiki	WFQd	Rf	ynf	nf	vgitvg	L14730
hum.ENG2	kneked	RR	pRta	F	tae	Q	lqr	L	kae	F	qa	nr	YIt	eq	RR	qt	LA	qe	LsLn	esqiki	WFQN	RR	aKi	KK	atgikn	L12699
hum.HLX1	tykrkr	sw	sRav	F	snl	Q	rkq	L	ekr	F	ei	qk	YVt	kp	DR	kq	LA	am	LgLt	daqvkv	WFQN	RR	mKw	RH	skeaqa	M60721
dro.BarH1	glskkq	RR	aRta	F	tdh	Q	lqt	L	eks	F	er	qk	YLS	vq	ER	qe	LA	hk	LdLs	dcqvkt	WFQN	RR	tKw	RR	qtavgl	M73079
hum.HEX	rplhkr	Kg	gqvr	F	snd	Q	tie	L	eke	F	et	qk	YLS	pp	ER	kr	LA	km	LqLs	erqvkt	WFQN	RR	aKw	RR	lkqenp	L16499
hum.TCL3A	rtppkk	KK	pRts	F	trl	Q	ice	L	ekr	F	hr	qk	YLa	sa	ER	aa	LA	ka	LkMt	daqvkt	WFQN	RR	tKw	RR	qtaeer	M75952
dro.Nk2	glpnkk	RR	rRvl	F	tka	Q	tye	L	err	F	rq	qr	YLS	ap	ER	eh	LA	sl	IrLt	ptqvki	WFQN	RR	yKt	RR	aqnekg	M27290
hum.NK3	vgggrk	KK	tRtv	F	srs	Q	vfg	L	est	F	dl	kr	YLS	ta	ER	ag	LA	as	LqLt	etqvki	WFQN	RR	nKw	RR	hvaael	M99587
cen.Ceh9	cqktrk	KK	aRtt	F	sgk	Q	vfe	L	ekq	F	ea	kk	YLS	ss	DR	se	LA	kr	LdVt	etqvki	WFQN	RR	tKw	KK	ieseke	X52811
hum.DLX7	aaakkl	RR	pRti	Y	ssl	Q	lqh	L	nqr	F	qh	tq	YLa	lp	ER	aq	LA	aq	LgLt	etqvki	WFQN	RR	sKy	KK	likqns	U73328
hum.GBX2	tstgkn	RR	rRta	F	tse	Q	lle	L	eke	F	hc	kk	YLS	lt	ER	sq	LA	ha	LkLs	etqvkv	WFQN	RR	aKw	RR	vkagna	U02080
hum.EMX1B	alarkp	RR	iRta	F	sps	Q	llr	L	eha	F	ek	nh	YVv	ga	ER	kq	LA	hs	LsLt	etqvkv	WFQN	RR	tKf	RR	qklee	X68880
chi.Ems1	ggpckm	RR	vRtv	F	kpe	Q	ler	L	ege	F	lk	qq	Ymv	gt	ER	vd	LA	at	LrLt	etqvkv	WFQN	RR	iKw	RR	qsmekq	U23054
chi.Prx1	kkkrkq	RR	nRtt	F	nss	Q	lqa	L	erv	F	er	th	YPd	af	vR	ed	LA	rr	VnLt	earvqv	WFQN	RR	aKf	RR	nermla	D13433
hum.CART1	vssskk	RR	hRtt	F	tsl	Q	lee	L	ekv	F	qk	th	YPd	vy	vR	eq	LA	lr	teLt	earvqv	WFQN	RR	aKw	RR	rerygq	U31986
cen.Ceh10	askrrk	RR	hRti	F	tqy	Q	ide	L	eka	F	qd	sh	YPd	iy	aR	ev	LA	gk	teLq	edriqv	WFQN	RR	aKw	RR	tektwg	X52812
xen.Xanf1	lswyrg	RR	pRta	F	trs	Q	iei	L	env	F	rv	ns	YPg	id	vR	ee	LA	sk	LaLd	edriqi	WFQN	RR	aKl	RR	shresq	X70282
mur.Lim3	eeaeata	RR	pRtt	i	tak	Q	let	L	ksa	Y	nt	sp	KPa	rh	vR	eq	LS	se	tgLd	mrqvqv	WFQN	RR	aKe	RR	lkkdag	L38248
mur.Lim1	nlgakr	Rg	pRtt	i	kak	Q	let	L	kaa	F	aa	tp	KPt	rh	iR	eq	LA	qe	tgLn	mrviqv	WFQN	RR	sKe	aR	mkqlsa	Z27410
hum.CUT	qpqhql	KK	pRvv	l	ape	E	kea	L	kra	Y	qq	kp	YPs	pk	ti	ed	LA	tq	LnLk	tstvin	WFHN	YR	sRi	RR	elfiee	M74099
hum.Ortoid1	lhccrk	RR	hRti	F	tde	Q	lea	L	enl	F	qe	tk	YPd	vg	tR	eq	LA	rk	VhLr	eekev	WFKN	RR	aKw	RR	qkrsss	A54677
cen.Unc30	gkiqkp	RR	gRth	F	tsh	Q	lte	L	enw	F	sr	nr	YPd	ma	cR	ee	IA	vw	IsLt	eprrv	WFKN	RR	aKw	RR	rernyv	L37867
hum.OTX1	atprkq	RR	eRtt	F	trs	Q	ldv	L	eal	F	ak	tr	YPd	if	mR	ee	VA	lk	InLp	esrvqv	WFKN	RR	aKc	Rq	qqqsgs	P32242
mur.Pax7	plkrkq	RR	sRtt	F	tae	Q	lee	L	eka	F	er	th	YPd	iy	tR	ee	LA	qr	tkLt	earfvq	WFSN	RR	aRw	RR	qaganq	U20792
hum.PAX3	plkrkq	RR	sRtt	F	tae	Q	lee	L	era	F	er	th	YPd	iy	tR	ee	LA	qr	akLt	earfvq	WFSN	RR	aRw	RR	qaganq	L07483
cen.Unc86	lpntdk	RR	kRts	i	aap	E	kre	L	eqf	F	kq	qp	RPs	ge	ri	as	IA	dr	LdLk	knvrv	WFKN	RR	qKq	RR	dfrrsqf	M22363
hum.BRN1B	aaqgrk	RR	kRts	i	evs	v	kga	L	esh	F	lk	cp	KPs	aq	ei	ts	LA	ds	LqLe	kevvrv	WFKN	RR	qKe	RR	mtppgg	Z11933
mur.Brn2a	ganerk	RR	rRtt	i	sva	a	kda	L	erh	F	ge	hs	KPs	sq	ei	mr	MA	ee	LnLe	kevvrv	WFKN	RR	qRe	RR	vktsln	X57512
dro.Pou28	silgrr	RR	kRts	i	ett	v	rtt	L	eka	F	lm	nc	KPt	se	ei	sq	LS	er	LnMd	kevirv	WFKN	RR	qKe	RR	inpsld	S80559
dro.Exd	ldarrk	RR	---n	F	skq	a	sei	L	ney	F	ys	np	YPs	ee	aK	ee	LA	rk	cgIt	vsqvs	WFQN	RR	iRy	KK	nigkaq	L19295
hum.PBX2B	ldarrk	RR	---n	F	skq	a	tei	L	ney	F	ys	np	YPs	ee	aK	ee	LA	kk	csIt	vsqvs	WFQN	RR	iRy	KK	nigkfq	X59841

Fig. 1. Main representatives of major evolutionary groups of homeodomains. All HOM aa sequences are presented according to a special format originally published in Ref. 6. At the top, three scales for numbering aa positions within HOM—outdated (2), standard (3), and new (6)—are shown in comparison. Also, boundaries of major HOM structural elements—Helix—Loop—Helix—Turn—Helix (HLHTH)—are outlined above the scales. The symbols * and ^ denote the most conservative aa positions throughout the HOM division and/or other large HOM categories such as Type, Branch, and Class. All amino acids involved in HOM group-specific patterns are capitalized and bold. The C-terminal segment of cen.Ceh13G5/L14730 probably containing cloning/sequencing errors is underlined. In the last column, GSDB ID, Accession Numbers of the corresponding nt/aa sequence entries are shown. Two major HOM groups, called type I and type II, are separated by the double-dashed line. Smaller HOM groups, named Branch and Class, are separated by single-dashed lines. Abbreviations used in HOM names: dro, *Drosophila melanogaster*; mur, *Mus musculus*; hum, *Homo sapiens*; cen, *Cenorhabditis elegans*; chi, *Gallus gallus*; and xen, *Xenopus laevis*.

Among the mentioned outer contacts, we should distinguish two major types: (i) *intrachain*, i.e., within a single polypeptide chain, which can be *local* (i.e., within a single structural element of a chain, e.g., within an α -helix, a β -sheet, etc.) or *distant*, i.e., between different structural elements within a single chain; and (ii) *interchain*, i.e., between different polypeptides in a multimeric protein. It is important to note that, generally, one and the same chemotype occurring in different regions of a protein may have different outer contacts and therefore different stereotypes. On the other hand, nonidentical but similar—*isofunctional*—chemotypes (defined later) occurring in the same region of a protein may still have the same stereotype. For instance, let us consider aa positions 37–40 in Fig. 1. The “LxLx” motif can be found at these positions within all of the type I HOMs. Given this, a few aa similar in their physical and chemical properties—in most S (serine), T (threonine), N (asparagine), and G (glycine)—can occupy the positions designated by “x.” Hence, we may suggest that all chemotypes like “LsLs,” “LsLt,” “LtLg,” etc., probably represent a single stereotype. The basis of this suggestion lies in the fact that aa positions 37–40 form the turn between helix II and helices III/IV, which is essential for the 3D configuration of this entire region of the HOM. Thus, the *location* of a chemotype along a protein chain is another factor determining the stereotype of a particular chemotype, in addition to its *primary structure*. Thus, while the term “position” is relevant only to a single aa, the term “location” is relevant to a chemotype. (Note: generally speaking, the stereotype–chemotype situation in structural biochemistry of proteins seems somewhat analogous to the phenotype–genotype correlations in genetics.)

We also found it helpful to distinguish six degrees of aa position conservativeness: unique, quasi-unique, strong, moderate, weak, and variable. In this context, an *aa position* within a chemotype location is called *unique* when *one and only one certain aa can occur at the given position, in a sample of homologous proteins*. In this case, the *C* index described above has a value of 1.0. On the contrary, a *position* is called *variable* when *any of the 20 common aa can occupy the given position*. In this case, *C* has a value close to 0.0. For instance, as follows from Fig. 1, only W (tryptophan) and N (asparagine) occur at positions 47 and 50, respectively, meaning that these aa positions are unique for this particular HOM sample. On the other hand, aa positions –6...–1 and 59–63 are variable. This might be expected because the positions are located outside of the HOM, being its N- and C-termi-

nal flanks. However, such variable positions also occur within HOMs, e.g., aa positions 9, 10, 20, 21, and 35, 36.

The other four degrees of aa position conservativeness represent different variants of preferences for the positions to be occupied by certain amino acids. For example, we call a position *quasi-unique* when *the position can be occupied predominantly by a certain aa and, occasionally, by another very similar one* (in structure and other properties). In this case, the *C* index has a value in limits of 0.99–0.91. Position 48 is one of such positions because almost all HOMs contain F (phenylalanine) at the position, but Y (tyrosine) can substitute F at this position in a few HOMs. It is important to note that F and Y are very similar in structure.

Further, we call a *position strongly conservative* when *two or three very similar aa can occur at the position with equal (or almost equal) probabilities*. In this case, values of *C* vary in limits of 0.90–0.70. Further, following the same logic, we call aa positions *moderately* or *weakly conservative* when values of their *C* indices vary in limits of 0.69–0.31 or 0.30–0.11, respectively.

It is important to emphasize that the degrees of conservativeness are not absolute. An aa position can be moderately or strongly conservative or even unique, depending on which particular HOM category we analyze. For instance, aa position 25 is moderately conservative throughout the entire HOM group because five aa can occur at this position. However, when we consider type I HOMs (see Fig. 1), four aa—all very similar—can be found at position 25. Moreover, when we consider type II HOMs, only one aa, P (proline), can occupy this position. Thus, position 25 is unique in the latter group, but at the same time, it is strongly conservative in the former one. We return to a discussion of the subject matter later, after an additional concept, “aa isofunctionality,” is introduced.

We can now define a *local functional motif* (FM) as a *chemotype in which some of its positions (at least a half) are quasi-unique or strongly conservative, while the others are moderately conservative or even variable*. Like any chemotype, a FM is involved in a certain stereotype. However, the latter should not essentially depend on a FM’s location along a protein. This suggestion seems likely because, in the evolution of homologous proteins, the conservativeness of a local FM is determined primarily by the conservativeness of its stereotype, if of course such a conservative chemotype is not merely an *evolutionary remnant*. We have already discussed an example of a local FM, in particular, “LxLx.” The following chemotypes—“KK,”

“RR,” “KR,” and “RK”—being N and C terminals of HOMs (see Ref. 6 and Fig. 1; aa positions 1,2 and 56,57) also represent a local FM that is discussed in detail later.

Finally, we can define a protein DP, considering that the concept of “domain” as such has been already introduced by other authors (8). A DP comprises *one or several local FMs intermittent by a certain number of variable aa positions strictly ordered according to their natural sequence in a particular protein*. In routine laboratory studies, a search for FMs and/or DPs is performed using multiple alignments of homologous proteins or their particular domains. Although there are computer programs designed to perform such alignments, global or local, precise algorithms to search for FMs and DPs do not yet exist. Our approach to the search was performed for the most part manually, and computer methods were used only for accessory purposes. This approach is based on the use of an extended concept of aa *isofunctionality* differentiated into several levels.

The aa Isofunctionality Concept

In general, this concept states that some aa play the same functional roles due to their similarity in structure and physical/chemical properties. This fact is thought to explain why *isofunctional* aa can mutually substitute each other without a loss of the protein function. Moreover, the fact underlies the genetic polymorphism of evolutionary related proteins. There are several indisputable cases of pairs of isofunctional aa such as D (aspartate) and E (glutamate), or K (lysine) and R (arginine), or even triplets such as V (valine), I (isoleucine), and L (leucine), etc. However, there is no agreement in the literature with respect to the exact number of isofunctional groups among all aa (9–15). Furthermore, there are discrepancies between isofunctional aa groupings based on comparisons of physical/chemical properties of aa and that based on their mutational substitutions. A thorough discussion of the subject matter is beyond the scope of this report and will be published elsewhere.

Here we present our approach to aa grouping based on both genetic data and physical/chemical properties. Three major levels of the aa isofunctionality—*strict*, *moderate*, and *weak*—are helpful in describing this kind of aa features. The suggestion follows from results of a statistical analysis of aa substitutions obtained from our updated HOM sample and gathered into a frequency matrix. The original version of the matrix was published earlier (6).

Each *ij* cell in this matrix, actually a 69×20 table, contains a number that is a value of the probability for each (*i*th) of the 69 aa positions studied within HOMs to be occupied by each (*j*th) of the 20 amino acids. Data collected in the matrix can be interpreted in two ways. We can calculate the *C* values for each aa position within the HOM and then identify unique, quasi-unique, conservative (strongly, moderately, or weakly), and variable positions. On the other hand, it is reasonable to consider, for example, only quasi-unique and strongly conservative positions and account only for those aa that occur at these positions with the highest probabilities. We can conclude that such aa are *strictly isofunctional genetically*. It turns out as a rule that such aa are also very similar in their physical/chemical properties. Likewise, if we now consider moderately conservative positions and account only for aa occurring at the positions with the highest probabilities, then such aa can be described as *moderately isofunctional genetically*. As a rule, such aa are relatively similar in their physical/chemical properties as well.

Based on the matrix and the approach described above, we generalized all data in the form of the diagram presented in Fig. 2. First, considering only chemical properties of aa, we divided all the 20 entities into three groups—hydrophobic, hydrophilic, and neutral ones—in concordance with standard classification. Then, considering only physical properties of aa, in particular, their molecular masses, all of them were divided into three other groups: small/light (55–105 Da), medium (105–135 Da), and large/heavy (135–185 Da). Finally, in each of the groups, we established subgroups with strictly and/or moderately isofunctional aa. As follows from the diagram in Fig. 2, we now have nine subgroups of aa isofunctionality. They are designated by Roman numerals. In routine computer analyses, however, it is more convenient to use traditional Arabic numerals to designate those groups. Such a numerical conversion of aa sequences usually presented in the one-letter code is important because it allows us to reduce the evolutionary noise considerably, i.e., to disregard accidental occurrences of weakly or even nonisofunctional aa at strongly or moderately conservative positions. Due to this, the gC_i index computed using probabilities of aa isofunctional groups is more sensitive in a search for conservative aa positions than the sC_i index computed using probabilities of individual aa. Furthermore, the presentation of homologous aa sequences using such numerical conversion essentially simplifies a search for FMs and DPs. This is demonstrated in the next section.

Fig. 2. Amino acid isofunctionality diagram based on chemical similarities and mutational substitutions of the 20 common amino acids in homeodomain-containing proteins (after Ref. 6; modified). The top bar subdivides the amino acids into three groups according to molecular masses of amino acids in daltons (MM in Da). The individual MMs are shown for each amino acid to the right of its symbol [s.a., S (serine), D (aspartate), E (glutamate), etc., in the standard one-letter code]. These MMs differ from the commonly published values for free amino acids each containing additional radicals H⁺ and OH⁻, i.e., the aa MMs given here are 18.02 Da less, which is the MM of one molecule of H₂O lost in a single peptide bond. At the bottom are the symbols and the names of chemical properties of the amino acids, including their major functional radicals. In the bottom right corner, the aa isofunctionality scale, ranging from 0 to 10 in conditional units, is presented. This scale is used in comparisons of aligned homologous aa sequences. It recognizes five levels—"identical" stands for the same amino acids occupying the same positions in two aa sequences compared (rank 10); "none" stands for an amino acid in a sequence that matches with a gap (rank 0) or with a non-isofunctional entity (rank 1) in another sequence. For explanation of the other three levels, see the text.

Evolutionary Classification of HOMs

Based on the data and the approach described above, we can now present several group-specific DPs that allow us to identify and establish major categories of the Evolutionary Systematics of Homeodomains (EvoSysHOMs). It seems obvious that the first, and the largest, of the categories should separate all known HOMs from all other protein domains. We call this category a **Division**. To be assigned to this group, an aa sequence should meet a set of criteria ("Recognizing Rules"; RRs), that we call the *HOM Division-specific pattern*. It is presented in Table I.

According to the pattern, first, hydrophilic positively charged aa—either **K** (lysine) or **R** (arginine)—occur predominantly at amino and carboxyl borders of

HOMs at positions 1 and 2, 51 and 52, 54, 56, and 57, with probabilities of $P = 0.90-0.91$, $0.90-0.99$, 0.99 , 1.0 , and 0.93 , respectively. Second, only hydrophobic (heterocyclic and aromatic) aa—**W** (tryptophan) and either **F** (phenylalanine) or **Y** (tyrosine)—can occupy positions 47 and 48, respectively. And, finally, only a hydrophilic aa, **N** (asparagine), can occupy position 50. This pattern allows one to distinguish any HOM from any unrelated protein domain with a probability greater than 0.90, except for noncanonical HOM varieties and uncommon polymorphic variants (including experimental errors; discussed later).

In actuality, it is not possible to gather all HOMs into a single group and, at the same time, completely

Table I. HOM Division-Specific Pattern^a

	Aa position													
	1	2	3–46	<u>47</u>	<u>48</u>	49	<u>50</u>	51	<u>52</u>	53	<u>54</u>	55	<u>56</u>	57
Letter pattern	R≡K	R≡K	Next 44 aa	W	F≡Y	x	N	R≡K	R	x	K≡R	x	K≡R	K≡R
Numeric pattern	5	5	Next 44 aa	7	8	x	3	5	5	x	5	x	5	5
Probability	0.91	0.90	...	1	1	·	1	0.90	0.99	·	0.99	·	0.99	0.93

^a (≡) "Strict aa isofunctionality" as shown in Fig. 2. The most conservative aa positions (unique, quasi-unique, and strongly conservative) are underlined. The symbol "x" stands for any amino acid at the position.

avoid indeterminacy with respect to whether (i) some true HOMs will be lost, whereas (ii) some definitely non-HOMs will be included in such a group. Our approach in establishing the HOM Division allows one to exclude the second possibility. However, the RR method cannot ensure that some true but noncanonical HOM aa sequences will not be lost. In fact, a few HOMs contain a 3-aa insertion in the *Loop* region and, at the same time, a 3-aa deletion in the N terminal (see Fig. 1; dro.Exd and hum.PBX2B; for review of this noncanonical HOM group, named *TALE*, see Ref. 16). There are also examples of larger insertions in other sites within the homeodomain as in the Prospero HOM in *Drosophila* or in the human and murine HNF1a,b HOMs. Such kind of atypical HOMs of course requires additional RRs to be included in the HOM division.

Using the same approach, we can now distinguish two major HOM types. The corresponding *Type-specific patterns* are presented in Table II. According to

the data, the following two RRs are effective. First, within **type I** HOMs, only three methyl-carrying and strictly isofunctional aa—I (isoleucine), **L** (leucine), and **V** (valine), or occasionally **M** (methionine), chemically similar to the former three—can occur at position 25. Second, within **type II** HOMs, only neutral **P** (proline) can occupy that position. In analyzing over 500 canonical HOMs, we did not find exceptions to these two rules. Such a type specificity of position 25 is not surprising since this position belongs to the *loop* between α-helix I and α-helix II of the HOM, which is thought to be essential for a precise conformation of helix III in its binding to relevant gene targets. It is also important to emphasize that type I encompasses all HOMs earlier assigned either to the *Caudal*–*Bithorax* and *Antennapedia* groups or to the *HLX*–*NKX*–*MSH* and *EMS*–*DLL* groups. At the same time, type II encompasses all HOMs earlier assigned to the "double domain-containing" HOM group including *LIM*–, *PAX*–, and *POU*–containing homeoproteins.

Table II. HOM Type-, Branch-, and Class-Specific Patterns^a

	aa position											
	1–2	3–23	<u>24</u>	<u>25</u>	26–28	<u>29</u>	30–46	47	48	<u>49</u>	50–57	
Type I HOMs: GBP	R≡K :5		Y≡F :8	V≡L≡I≡M :9		x		W :7	F≡Y :8	Q :4	See Tab I	
Branch I.A (APA)	R≡K :5		Y≡F :8	V≡L≡I≡M :9		R≡K :5		W :7	F :8	Q :4	See Tab I	
Branch I.B (EMI)	R≡K :5		Y :8	V≡L≡I≡M :9		D≡E :4		W :7	F=Y :8	Q :4	See Tab I	
Type II HOMs: NED	R≡K :5		x	P :1		x		W :7	F :8	x	See Tab I	
Branch II.A	R≡K :5		x	P :1		x		W :7	F :8	Q :4	See Tab I	
Class II.A1 (PHA)	R≡K :5		Y :8	P :1		x		W :7	F :8	Q :4	See Tab I	
Class II.A2 (LIM)	R≡K :5		K≡R/N :5/3	P :1		x		W :7	F :8	Q :4	See Tab I	
Branch II.B	R≡K :5		x	P :1		x		W :7	F :8	x	See Tab I	
Class II.B1 (OTD)	R≡K :5		Y :8	P :1		x		W :7	F :8	K≡R/H :5/6	See Tab I	
Class II.B2 (PAX)	R≡K :5		Y :8	P :1		x		W :7	F :8	S :3	See Tab I	
Class II.B3 (POU)	R≡K :5		K≡R :5	P :1		x		W :7	F :8	C :2	See Tab I	
Class II.B4 (PEX)	R :5		Y :8	P :1		x		W :7	F :8	G :2	See Tab I	

^a The aa positions essential in assigning HOM Type (25) Branch (29, 49) and Class (24, 49)-specific patterns are underlined. Numerals next to amino acids represent the corresponding aa isofunctional groups (see Fig. 2). The symbol "x" stands for any amino acid at the position. Skipped aa positions are denoted by "...". Abbreviations used in naming HOM categories: *GPB*, general body plan; *APA*, anteroposterior axis; *EMI*, epithelial–mesenchymal interactions; *NED*, neuroectodermal derivatives; *PHA*, PHOX–aristaless families; *LIM*, LIM and HOM double-domain containing; *OTD*, ortoid–ortodenticle families; *PAX*, paired and HOM double-domain containing; *POU*, POU and HOM double-domain containing; *PEX*, PBX/extradenticle family.

Further, we introduce a new category—*Branch*—that allows us to divide all type I HOMs into branches **I.A** and **I.B**. The key distinction between these two groups is determined by aa occurring at position 29. In fact, each of the **I.A** HOMs contains only basic positively charged aa—either **R** (arginine) or **K** (lysine)—at that position, whereas each of the **I.B** HOMs contains there only acidic negatively charged aa—either **D** (aspartate) or **E** (glutamate). These rules indeed establish *Branch-specific patterns* that are valid only for the type I HOMs. Considering the well-known expression patterns of the proteins, it is important to emphasize this strong correlation established between such a structurally based HOM grouping and those expression patterns. In fact, branch **I.A** encompasses mainly *homeotic selector* genes and a few *pair-rule/segment polarity* genes as well, all of which are transmitters of very early positional information and, thus, are involved directly in the general body plan formation, specifying the anteroposterior axis of the developing embryo. On the other hand, branch **I.B** encompasses mainly those HOMs that function later and control particular epithelial–mesenchymal interactions, cellular differentiation, and organogenesis.

Unlike type I HOMs, the essential distinctions among particular categories within type II HOMs are determined by aa occurring at positions 24 and 49. Namely, while only **Q** (glutamine) occurs at position 49 within any of the type I HOMs, at least seven aa from different aa isofunctional groups can occur at that position within type II HOMs. Moreover, these seven allow us to recognize distinctive HOM categories in the latter type: branches and a new category called *Class*. For instance, all the type II HOMs containing **Q** (glutamine) at position 49 can be grouped into branch **II.A**. Within this branch we establish two classes, **II.A1** (“Phox-Aristaless”; PHA) and **II.A2** (“LIM”), which can be distinguished from each other by aa at position 24 (see Table II). On the other hand, all other type II HOMs, which contain, at position 49, any of the following six aa—**R** (arginine), **H** (histidine), **K** (lysine), **S** (serine), **C** (cysteine), or **G** (glycine)—can be grouped into branch **II.B**. Within this category, **K** (Lys) at the position identifies Class **II.B1**, “Ortoid–Ortodenticle”; **S** (Ser) identifies Class **II.B2**, “PAX”; **C** (Cys) identifies Class **II.B3**, “POU”; and **G** (Gly) identifies Class **II.B4**, “PEX” (PBX/Extradenticle). Interestingly, the majority of the type II HOMs is involved primarily in the development of neuroectodermal derivatives.

A more complicated task is to define distinctive HOM *Families* and to recognize their *Family-specific patterns*. We found a short aa segment encompassing

positions 24–30 to be the most effective for this purpose. It is indeed highly conservative within any particular family, though varying from one family to another. Likewise, other short segments within the HOM, and even particular sites, turn out to be helpful in recognizing HOM category smaller ranks such as distinctive *Varieties* and *Sorts*. Formal definitions and a complete discussion of all proposed evolutionary categories, as well as of some other related subjects, will be published elsewhere.

One may ask, Why do we need so many HOM categories? and, What is the difference between groups such as Type, Branch, Class, Family, etc.? The answer is simple: we encounter too great a diversity among HOMs belonging to a wide range of animal proteomes. Given this, some HOMs are more similar and some are less similar to each other. It seems natural that larger categories, such as Type, Branch, and Class, should show a higher level of HOM diversity, while smaller categories, Family, Variety, and Sort, encompass much more similar HOMs. Thus, criteria determining the categories are based on degrees of HOM diversity, and strictly speaking, the degrees can be formally defined and estimated quantitatively. On the other hand, to assign particular HOMs to particular categories, we need to find category-specific patterns. Contrary to the criteria of HOM diversity, such patterns should recognize, first of all, similarities among individual HOMs. Thus, two sets of criteria outfitting a satisfactory evaluation both of HOM diversity and of HOM similarity are mutually correlated but not identical.

Further, it is reasonable to mention how the proposed EvoSysHOMs can be an effective tool in distinguishing genetic polymorphism of HOMs from a variety of experimental errors that are sufficiently common that they cannot be disregarded. For example, among the *Cenorhabditis elegans* HOMs submitted to GenBank there are two aa sequences that can be certainly assigned to the *Engrailed*-like family: Ceh12 (gnb:X17076) and Ceh13G5 (gnb:L14730). Comparing the carboxyl terminals of these two HOMs with those of the other members of the family at aa positions 47–57, we can see that the relevant segments are as follows.

In Ceh13G5;	“WFQDRFYNFNF”	(or “78445883838”)
In Ceh12;	“WFQNRMRMKNKR”	(or “78435595355”)
In all others;	“WFQNKRAKxKK”	(or “78435525x55”)

Now it seems obvious that the segment “DRFYNFNF” in Ceh13G5 (see Fig. 1) is a cloning/sequencing artifact because position 50 is *unique* and it can be occupied only by N (Asn), according to the

HOM Division-specific pattern earlier discussed. The remaining aa segment in Ceh13G5 also does not fit the Engrailed-like family-specific pattern well whereas the relevant segment in Ceh12 fits this pattern completely.

One more example relates to the *Caudal*-like family. According to results of our global alignments of the entire HOM family (7), there are four Cad/Cdx *Varieties* in Mammals; Cdx1, -2, -3, and -4. All these, as well as several other orthologous and paralogous Caudal-like HOMs from different vertebrates, have chemotypes "MYV" (Met-Tyr-Val), "MYG" (Met-Tyr-Gly), or "MY-" (Met-Tyr-gap), at their amino terminals. Among the mentioned HOMs, the murine Cdx1 is represented by four separate entries, in GSDBs: pir:A49303, swp:P18111, gnb:M37163, and gnb:M80463 (without counting additional six quasi-copies of the documents, created by the NCBI Database). Interestingly, all these entries refer to the same two publications (17,18). However, only A49303 and P18111 present the correct Cdx1 aa sequence beginning with "MYV" published in Ref. 18. The other two entries (gnb:M37163/M80463) present a much shorter aa sequence beginning with "MAR" (Met-Ala-Arg). As our analysis showed, the two gnb:entries contain complex semantic errors in the nucleotide sequences submitted: (i) in M37163 and Ref. 17, a dinucleotide "cc" is missing at positions 214–215; (ii) this dinucleotide is present in Ref. 18 but it is missing in M80463, although the entry refers to Ref. 18; (iii) moreover, an incorrect translation start is indicated in the above entry, contrary to Ref. 18. Thus, the "MAR" amino-terminal Cdx1 is not a distinctive genetic HOM variety or even an alternatively splicing variant, but a result of experimental errors in Ref. 17 and semantic errors committed in the computer processing of gnb:M80463. Meanwhile, the discovery of the errors became possible due to the evolutionary analysis of HOMs described here. Two examples discussed above are not unique, nor are they even the most extreme among GSDB entries.

Finally, we suggest that major applications of the Evolutionary HOM Classification are related to the prediction of yet uncovered orthologous HOM Varieties and Sorts, as well as to a deeper understanding of evolutionary and developmental mechanisms of HOM protein diversity.

ACKNOWLEDGMENTS

The authors are grateful to Ch. Strom for general advice and stimulating critical review of the results

and the manuscript. We thank Alexander Zhadanov for invaluable informatics support. We also thank Alex Kister, with whom we discussed some conceptual aspects of the study.

REFERENCES

- Gindilis VM, Ageev SV: HuNeG project: The progress report of 1993-94 on the Scottish Rite SRP Huneg grant award, Oct 1994, pp 1–31. (Office of the SRP, P.O. Box 519, 33 Marret Road, Lexington, MA 02173)
- Scott MP, Tamkun JW, Hartzell GW: The structure and function of the homeodomain. *Biochim Biophys Acta* 1989;989: 25–48
- Burglin TR: A comprehensive classification of homeobox genes: *In* Guidebook to the Homeobox Genes, D Duboule (ed). New York, Oxford University Press, 1994, pp 25–71
- Stein S, Fritsch R, Lemaire L, Kessel M: Checklist: Vertebrate homeobox genes. *Mech Dev* 1996;55:91–108
- Abouhief E, Akam M, Dickinson WJ, Holland PWH, Meyer A, Patel NH, Raff RA, Roth Wray GA: Homology and developmental gene. *Trends Genet* 1997;13:432–433
- Gindilis V, Banikazemi M, Vyasankin A, Verlinsky O, Matveyev I, Verlinsky Y: Borders, patterns, and distinctive families of homeodomains. *J Assist Reprod Genet* 1994;11:244–269
- Gindilis VM, Verlinsky Y: Reconstruction of mutational history in the evolution of the caudal-like homeodomain family. *J Mol Evol* (in press)
- Islam SA, Luo J, Sternberg MJE: Identification and analysis of domains in proteins. *Prot Eng* 1995;8:513–525
- Schmidt W: Phylogeny reconstruction for protein sequences based on amino acid properties. *J Mol Evol* 1995;41: 522–530
- Jiwani NG, Liebman MN: Structure-function analysis of amino acid substitutions in proteins: *In* Molecular Modeling, TF Kumosinski, MN Liebman (eds). Washington, DC, Am Chem Soc, 1994, pp 184–206
- Bowie JU, Luthy R, Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;253:164–170
- Taylor WR: The classification of amino acid conservation. *J Theor Biol* 1986;119:205–218
- Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA: Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Prot Chem* 1985;4:23–55
- Ratner V, Jarkich A, Kolchanov N: *In* Problems of the Molecular Evolution Theory. Novosibirsk, Nauka, 1985, pp 63–74 (in Russian)
- Swanson R: A unifying concept for the amino acid code. *Bull Math Biol* 1984;46:187–203
- Burglin TR: Analysis of TALE superclass homeobox genes. *Nucleic Acids Res* 1997; 25:4173–4180
- Duprey P, Chowdhury K, Dressler GR, Balling R, Simon D, Guenet J-L, Gruss P: A mouse gene homologous to the *Drosophila* gene *caudal* is expressed in epithelial cells from the embryonic intestine. *Genes Dev* 1988;2:1647–1654
- Hu Y, Kazenwadel J, James R: Isolation and characterization of the murine homeobox gene *Cdx-1*. *J Biol Chem* 1993; 268:27214–27225