Protein Engineering vol.12 no.11 pp.943–951, 1999

Analysis of heregulin symmetry by weighted evolutionary tracing

Ralf Landgraf, Daniel Fischer[1] and David Eisenberg[2]

University of California, UCLA–DOE Laboratory of Structural Biology and Molecular Medicine and Molecular Biology Institute, 405 Hilgard Avenue, Box 951570, Los Angeles, CA 90095-1570, USA
[1]Present address: Department of Math and Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

[2]To whom correspondence should be addressed

Heregulins are members of the protein family of EGF-like growth and differentiation factors. The primary cell-surface targets of heregulins are heterodimers of the EGF-receptor homolog HER2 with either HER3 or HER4. We used a weighted evolutionary trace analysis to identify structural features that distinguish the EGF-like domain (hrg) of heregulins from other members of the EGF family. In this analysis, each amino acid sequence is weighted according to its uniqueness and the variability in each position is assigned by an amino acid substitution matrix. Conserved residues in heregulin that are variable in other EGF-like domains are considered possible specificity-conferring residues. This analysis identifies two clusters of residues at the foot of the boot-shaped hrg domain. The residues in one cluster are recruited from the N-terminus; those in the other are from the Ω-loop region and show a weak sequence similarity to the N-terminal residues at the opposite side of the boot. The remaining residues with high conservation scores distribute themselves into these two distinct surfaces on hrg. This pseudo-twofold symmetry and the presence of two distinct interfaces may reflect the preference of hrg for heterodimeric versus homodimeric HER complexes.

Keywords: evolutionary trace/HER/heregulin/receptor heterodimer/symmetry

Introduction

Receptor tyrosine kinases play a crucial role in a wide variety of cellular signaling events. The EGF receptor family includes in addition to the EGF receptor (EGFR) at least three homologs, HER2, 3 and 4 (also named cErbB 2, 3, 4). HER2, 3 and 4 will be collectively referred to as HERs (Human Epidermal growth factor Receptors). HERs are involved in cell proliferation and neuronal development (Gassmann *et al*., 1995; Lee *et al*., 1995). HER2 is overexpressed in 25–30% of breast carcinoma in humans and overexpression has been linked to a poorer prognosis (Slamon *et al*., 1989). Overexpressed HER3 or HER4 has likewise been implicated in cellular transformation (Earp *et al*., 1995).

Ligands have been identified that stimulate the formation of EGFR dimers and the subsequent tyrosine phosphorylation of EGFR. The identification of ligands for HERs has been complicated by the fact that only HER4, the least abundant receptor in this family, is believed to be a fully functional receptor by itself (Plowman *et al*., 1993a,b). HER3 binds heregulins with high affinity but lacks tyrosine kinase activity (Guy *et al*., 1994). HER2, the most abundant member of the family, exhibits a potent kinase activity but no ligands have yet been identified that bind to or activate HER2 directly. HER2 may be constitutively active when it is overexpressed at high levels on the cell surface (Pegram *et al*., 1997) and its potent tyrosine kinase activity can be stimulated by the association with other members of this receptor family (Karunagaran *et al*., 1996; Alimandi *et al*., 1997). The dimer generated by HER3 and HER2 binds heregulin with high affinity (Sliwkowski *et al*., 1994) and appears to be the dominant means of HER2 activation (Tzahar *et al*., 1994; Karunagaran, *et al*., 1996).

The known ligands for HERs are collectively named heregulins but are also known by other terms including acetylcholine receptor inducing activity (ARIA), neu differentiation factor (NDF), glial growth factor (GGF) and neuregulin. More than 10 isoforms are expressed in a variety of tissues and multiple isoforms (α, β1, β2, β3, γ) are derived from a single gene by alternative splicing (Marchionni *et al*., 1993). Heregulins are either secreted directly as a soluble species or are derived from a larger, presumably membrane-bound, precursor by proteolytic cleavage (Holmes *et al*., 1992; Marchionni *et al*., 1993). The ability of these ~240 residue proteins to bind to cognate receptors and elicit tyrosine phosphorylation resides almost exclusively in the ~60 residue EGF-like domain.

The recombinant, 60 amino acid EGF-like domain (hrg) of heregulin has been used to target HERs with liposomes (Park *et al*., 1995), directed toxins (Kihara and Pastan, 1995; Siegall *et al*., 1995; Landgraf *et al*., 1998) or viral DNA delivery systems (Han *et al*., 1995). The targeting characteristics of those constructs demonstrate that heterodimers of HER2/HER3 and to a lesser extent HER2/HER4 are the primary target on the cell surface. A comparison of the NMR structure of the 63 amino acid EGF-like domain of heregulin α (Jacobsen *et al*., 1996) with the corresponding domain of EGF (Kohda and Inagaki, 1992) shows high structural similarity in regions with well defined secondary structure [root mean square difference = 1.4 Å (Jacobsen *et al*., 1996)]. The common structural features in these ligands are three strands of β-sheet, stabilized by three conserved disulfide bridges. The most noteworthy differences are in the N-terminus (residues 1–13), the disordered C-terminus (residues 50–63) and the so-called Ω-loop (residues 24–31), which contains a three residue insertion, absent in EGF.

Information on those residues important for the interaction between hrg and its receptor comes predominantly from two sources. Barbacci *et al*. (1995) generated an extensive set of peptides in which the various inter-cysteine regions of EGF and hrg were exchanged. Jones *et al*. (1998) carried out a complete alanine scan of hrg β. The results from both sets of experiments indicate that the N-terminus of hrg is important in conferring receptor specificity. A chimeric EGF-peptide carrying the N-terminal five amino acids of hrg (Barbacci *et al*., 1995) binds both EGFR and HER2/HER3 dimers with comparable affinity (approximately 5–10 nM) (Tzahar *et al*.,

1997). Alanine scanning further indicates that many residue positions critical for binding are shared between EGF and hrg.

When a family of sequences is available, the identification of functionally important residues may be approached by computational means. Different methods have been proposed for such an analysis. One approach introduced by Casari *et al*. (1995) uses a vectorial analysis of sequence profiles to identify functionally important residues within a protein. Another approach (Lichtarge *et al*., 1996), called evolutionary tracing, also starts with a multiple sequence alignment of a protein family. The degree of conservation in each position is then calculated for different subsets of aligned sequences. The resulting score is visualized on the three-dimensional structure of one member of the family to identify functionally important clusters of residues.

This evolutionary trace analysis assumes structural conservation of the protein family under investigation. For the EGF-like family of growth factor ligands, this requirement is fulfilled. However, the characterization of solvent-exposed and buried residues, used in the original evolutionary trace analysis (Lichtarge *et al*., 1996) to discriminate between structurally or functionally important residues, has little discriminating power in the case of the EGF-like domain: in this small domain most residues are surface residues. To add discriminating power to our analysis, we introduced a quantitative measure of residue variation at each position. Using this modified evolutionary trace method, we analyzed a set of 45 EGF-like domains found in growth factors of various species. We compared the sequences for the heregulin-like subset of ligands with the larger family of EGF-like ligands. Our goal was to identify specificity conferring positions within heregulin.

Materials and methods

Partitioning of sequences and cluster assignment

Forty-five sequences for the EGF domain found in growth factors of various species were compiled following a BLAST search with the sequence of human EGF and heregulin alpha. Omitted from the analysis were artificial growth factor sequences, non-human sequences that did not contribute additional sequence diversity and sequences with probability scores above 10^{-5}. The sequence of the EGF-like domain of heregulin

a

```
                          1         10        20        30        40        50        60
                          |         |         |         |         |         |         |
Neuregulin 2_(human)      GHARKCNETAKSYCVNGGVCYYIEGINQLS___CKCPVGYTGDRCQQFAMVNFSKHLGFELKE
NTAK alpha 2b (rat)       GHAR_CNETAKSYCVNGGVCYYIEGINQLS___CKCPNGFFGQRCLEKLPLRLYMPAEELYQK
NTAK (rat)                GHARKCNETAKSYCVNGGVCYYIEGINQLS___CKCPNGFFGQRCLEKLPLELK__AEELYQK
NTAK alpha 2 (rat)        GHARKCNETAKSYCVNGGVCYYIEGINQLS___CKCPNGFFGQRCLEKLPLKQK__AEELYQK
NTAK (human)              GHARKCNETAKSYCVNGGVCYYIEGINQLS___CKCPNGFFGQRCLEKLPLKQK__AEELYQK
NDF (human)               SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCQPGFTGARCTENVPMKVQNQA_______
Heregulin alpha (human)   SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCQPGFTGARCTENVPMKVQNQAEELYQK
Heregulin alpha 2 (human) SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCQPQFTGARCTENVPMK____________
NDF (rat)                 SHLIKCAEKEKTFCVNGGECFTVKDLSNPSRYLCKCQPGFTGARCTENVPMKVQTQAEELYQK
Heregulin beta 3 (human)  SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVMASFYST________
GGF (Bos taurus)          SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVMASFYST________
GGF (human)               SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVMASFYSTSLPE___
Heregulin gamma (human)   SHLVKCAEKEKTFCVNGGECFMVKDLSNPS____CKCPNEFTGDRCQNYVMASFYSTSLPE___
Neuregulin beta 1a (human) SHLTKCDIKQKAFCVNGGECYMVKDLPNPPRYLCRCPNEFTGDRCQNYVEAEELYQ________
ARIA (chicken)            SHLTKCDIKQKAFCVNGGECYMVKDLPNPPRYLCRCPNEFTGDRCQNYVEAEELYQITGICIA
Neuregulin beta 2a (chicken) SHLTKCDIKQKAFCVNGGECYMVKDLPNPPRYLCRCPNEFTGDRCQNYVKAEELYQ________
Heregulin beta1 (human)   SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVEAEELYQ________
NDF (human)               SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVEAEELYQ________
Heregulin beta 2 (human)  SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVKAEELYQ________
NDF (human)               SHLVKCAEKEKTFCVNGGECFMVKDLSNPSRYLCKCPNEFTGDRCQNYVKAEELYQ________
Neuregulin alpha 1(Xenopus) GHLIKCSDKEKTYCVNGGECYVLNGITSSNQFMCKCPGFTGARCTETDPLRVVRSAEELYQK
Neuregulin 3 (mouse)      EHFKPCRDKDLAYCLNDGECFVIETLTGSHKH_CRCKEGYQGVRCDQFLPKTDSILHLGIE__
Epiregulin (human)        VSITKCSSDMNGYCLH_GQCIYLVDMSQNY____CRCEVGYTGVRCEHFFLTVHQPLALT_____
Epiregulin (rat)          VQITKCSSDMDGYCLH_GQCIYLVDMREKF____CRCEVGYTGLRCEHFFL_____________
HBEGF (rat)               KKRDPCLKKYKDYCIH_GECRYLKELRIPS____CHCLPGYHGQRCHGLTLPVENPLTTVL___
HBEGF (mouse)             KKRDPCLRKYKDYCIH_GECRYLQEFRTPS____CKCLPGYHGHRCHGLTLPVENPLTTVL___
HBEGF (pig)               KKRDPCLRKYKDFCIH_GECKYVKELRAPS____CICHPGYHGERCHGLSLPVKNRLTTIL___
SDGF (hamster)            RKKNPCDTEFQNFCIH_GECKYIENLEAVT____CNCHLDYFGERCGEKSMKTHSGDKIALAAI
SDGF (rat)                KKKNPCAAKFQNFCIH_GECRYIENLEVVT____CHCHQDYFGERCGEKTMKTQKKDKIALAA_
Amphiregulin (human)      KKKNPCNAEFQNFCIH_GECKYIEHLEAVT____CKCQQEYFGERCGEKSMKTHSMIKIALAAI
Amphiregulin (sheep)      RRKNLCDTEFQNFCIH_GKCTFLEQLETVS____CQCYPEYFGERCGEK_____________
EGF (human)               NSDSECPLSHDGYCLHDGVCMYIEALDKYA____CNCVVGYIGERCQYRDLKWWELRQK_____
EGF (pig)                 NSYSECPPSHDGYCLHGGVCMYIEAVDSYA____CNCVFGYVGERCQHRDLKWWELR_______
EGF (rat)                 NSNTGCPPSYDGYCLNGGVCMYVESVDRYV____CNCVIGYIGERCQHRDLRWWKLRQR_____
EGF (mouse)               NSYPGCPSSYDGYCLNGGVCMHIESLDSYT____CNCVIGYSGDRCQTRDLRWWELRQK_____
EGF (horse)               NSYQECSQSYDGYCLHGGKCVYLVQVDTHA____CNCVVGYVGERCQHQDLR____________
TGF alpha (rat)           SHFNKCPDSHTQYCFH_GTCRFLVQEEKPA____CVCHSGYVGVRCEHADLLA___________
TGF alpha (mouse)         SHFNKCPDSHTQYCFH_GTCRFLVQEEKPA____CVCHSGYVGVRCEHADLLA___________
TGF alpha (human)         SHFNDCPDSHTQFCFH_GTCRFLVQEDKPA____CVCHSGYVGARCEHADLLA___________
TGF alpha (pig)           SHFNDCPDSHSQFCFH_GTCRFLVQEDKPA____CVCHSGYVGARCEHADLLA___________
TGF alpha (bovine)        SHFNDCPDSHSQFCFH_GTCRFLVQEEKPA____CVCHSGYVGARCEHADLLA___________
TGF alpha (hamster)       SHFNECPDSHTQFCFH_GTCRFLVQEEKPA____CVCHSGYVGARCEHADLLA___________
TGF alpha (rabbit)        SHFNQCPDSHTQFCFH_GTCRFLVQEDKPA____CVCHSGYVGARCEHADLLA___________
Betacellulin (human)      GHFSRCPKQYKHYCIK_GRCRFVVAEQTPS____CVCDEGYIGARCERVDLFYLRGDVICL___
Betacellulin (mouse)      THFSRCPKQYKHYCIH_GRCRFVVDEQTPS____CICEKGYFGARCERVDLFYLQQDVVCL___
```

alpha, as represented in the NMR structures of the latter, was used to define the length of the sequence alignment. The multiple sequence alignment program PILEUP (Devereux *et al*., 1984) of the GCG sequence analysis package was used to generate the multiple sequence alignment and dendrogram. The default settings were used for the analysis.

Weighted evolutionary trace analysis

The method of Lichtarge *et al*. (1996) was modified to incorporate a quantitative description of residue variability at each position, an advantage when dealing with highly homologous sequences. To identify the variability in each position p of the multiple alignment M, created by PILEUP, we computed a variability measure V_p as follows:

$$V_p = \sum_{i=1}^{n} \frac{v_{i,p}}{n} \tag{1}$$

where n is the number of sequences in the multiple alignment and $v_{i,p}$ is the variability measure at position p, using sequence i as a reference (V_p is the average of the $v_{i,p}$s). To avoid over-representation of identical or very similar sequences, each sequence was weighted (W_i) according to its level of similarity using the method described by Sibbald and Argos (1990). The sum of all weights equals 100%. The individual variability measures $v_{i,p}$ were computed as follows:

$$v_{i,p} = \sum_{j\neq i,\ j=1}^{n} \frac{S_{M(i,p),M(i,p)} - S_{M(i,p),M(j,p)}}{S_{M(i,p),M(i,p)}} \times W_i \tag{2}$$

where $S_{M(i,p),M(j,p)}$ is the value of substituting the amino acid at position p in sequence i with the amino acid in the same position, but of sequence j, as given by a positive Gonnet substitution matrix (Benner *et al*., 1994). Gaps in the multiple sequence alignment were assigned the maximum substitution penalty. Several calculations with alternative gap penalties did not change the outcome significantly. Note that V_p yields values from zero to one, with zero representing no variability. This overall variability score for each position (V_p) was calculated for all the EGF-like ligands in Figure 1 [V(all)] and for the various heregulin clusters defined by partition I–IV [V(hrg)].

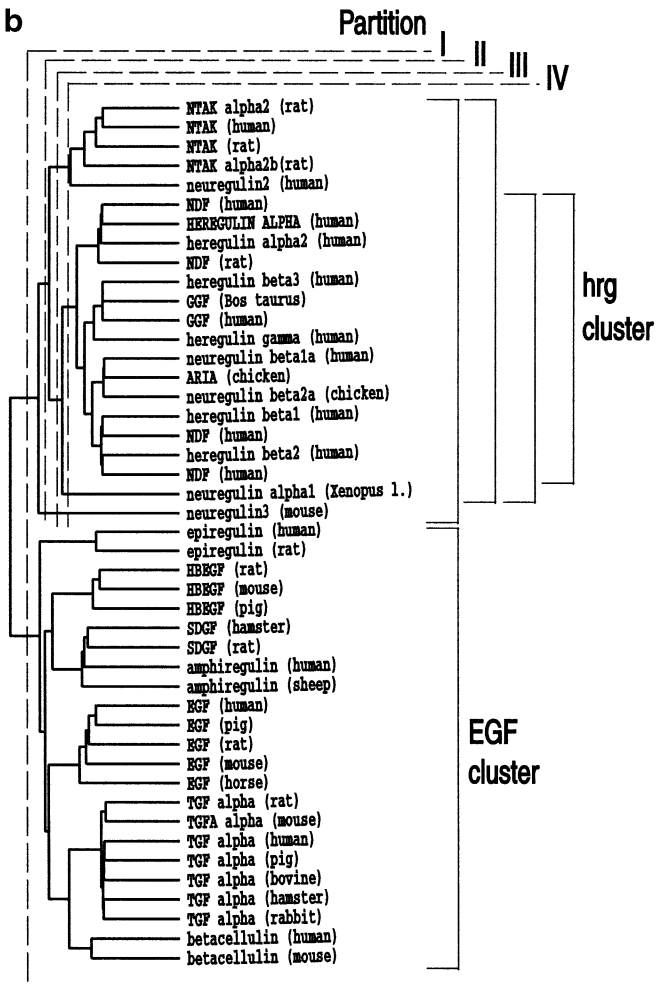

Fig. 1. Multiple sequence alignment and dendrogram of sequences used in this analysis. (**A**) The multiple sequence alignment of the 63 amino acid EGF-like domain of heregulin alpha and the corresponding segments of homologous EGF-like growth factor domains. These show that despite high sequence conservation within groups of ligands, only nine (highlighted) positions are fully conserved in all 45 sequences. (**B**) A dendrogram separates the 45 sequences into two main clusters that correlate with preferential binding to EGFR (EGF cluster) or HER (hrg cluster). The hrg cluster was further subdivided into partitions as indicated by dashed lines. Those sequences remaining in the same cluster as the reference sequence [HEREGULIN ALPHA (human)] were compared by weighted evolutionary tracing against the complete set of sequences by computing heregulin specific conservation scores ($\boldsymbol{C}_{\mathbf{hrg}}$).

Results

Partitioning of sequences and positional variability

Forty-five sequences of EGF-like domains were selected for the evolutionary trace analysis. Figure 1a shows those positions for each sequence which align to the reference sequence, the EGF-like domain of heregulin alpha. A dendrogram of those 45 sequences (Figure 1b) reflects the functional differentiation of EGF-like ligands: two main branches separate ligands that primarily act on EGFR from ligands that target HERs. Most smaller subdivisions place the homologs from different species together. The most remarkable exception is neuregulin 3 from mouse, which only clusters with the hrg-type ligands at the broadest possible partition. Its association with the hrg family is even weaker than that of the *Xenopus* homolog of neuregulin alpha.

We introduced a total of four partitions into the dendrogram that further subdivide the hrg-type ligands according to their level of similarity. The first partition subdivides a cluster centered around EGF from heregulin-like ligands. Partition II separates the distant mouse homolog, neuregulin 3, from the heregulin cluster. Mouse neuregulin 3 is thought to activate HER4 preferentially (Zhang *et al*., 1997). A major subdivision is introduced with partition III, which removes NTAKs (neural and thymus derived activators of ErbB kinases). NTAKs are believed to bind primarily to HER3 and HER4 and to activate both HER2 and EGFR via heterodimerization with HER3 and HER4 (Higashiyama *et al*., 1997). Finally, at partition IV only the most highly conserved members of the heregulin family remain clustered together. With few exceptions, those sequences represent homologs or splicing isoforms of heregulin with almost identical EGF-like domains.

The evolutionary trace analysis used here starts with the calculation of a positional variability score (V_p), as defined by Equations 1 and 2. The heregulins in partition IV share a very high degree of sequence conservation, ranging from 68 to 85%. For the calculation of variability scores, all sequences were weighted according to sequence similarity. This was done to ensure that almost identical sequences do not dominate the analysis while ensuring that the information present in small substitutions or variations in size is harvested. Identical sequences differ outside the 63 amino acid window of analysis and, owing to the weighting scheme, the number of copies used in the analysis has no effect on the outcome of the calculation. These duplicate entries were kept merely to give a representation of the level of conservation within a group of ligands. The positional variability scores for the entire data set (partition 0) show a marked increase past position 51, a methionine in heregulin alpha (Figure 2). A comparison of the variability at partition 0 with the variability of the two main clusters in partition I shows that the majority of the positional variability is contributed by sequences in the EGFR cluster. However, despite the high level of sequence conservation within the hrg cluster, only 12 positions are fully conserved, six of which are the cysteines that form the three conserved disulfide bonds (highlighted in dark gray in Figure 1a). Of the remaining six positions (His2, Asn6, Gly18, Phe40, Gly42 and Arg44), half (Gly18, Gly42, Arg44, highlighted in light gray in Figure 1a) are conserved in all sequences.

Analysis of heregulin specific conservation

To identify those residues most likely to be involved in conferring specificity towards HERs, we calculated for each position p the ratio of variability for all EGF-like ligands

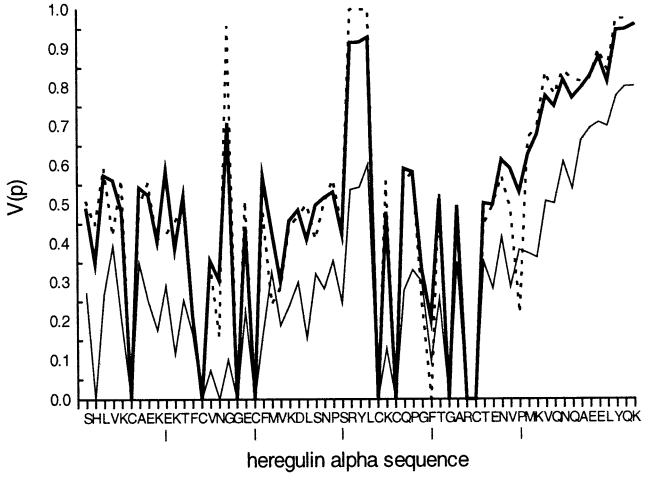

Fig. 2. Positional variability [$V(p)$] scores with respect to the heregulin alpha reference sequence. The variability scores for each position were calculated from the complete data set (heavy black line) and the EGF (dashed line) and hrg cluster (thin black line) alone. The EGF cluster contributes more to the overall variability, but both clusters show a higher conservation within the first 50 positions, known to be sufficient for the biological function of most EGF-like ligands.

[$V_{all}(p)$] to the variability within the hrg cluster [$V_{hrg}(p)$] at the different partitions indicated in Figure 1. We express this variability ratio as hrg specific conservation [$C_{hrg}(p)$]:

$$C_{hrg}(p) = \frac{V_{hrg}(p) + 1}{V_{all}(p) + 1} - 1 \qquad (3)$$

$C_{hrg}(p)$ can vary between $+1$ and $-\frac{1}{2}$ and is designed to assign high scores to positions that display overall variability but relative conservation within the cluster of interest, in this case heregulins. Positions with high overall variability or conservation are both downweighted. Figure 3 shows the resulting specificity scores for the EGF-like domain of heregulin alpha calculated for the hrg cluster at partition I–IV. As the partition number increases, the size and variability of the heregulin cluster decreases. As a result, positional conservation scores increase.

Above what threshold is the value of C_{hrg} meaningful? To estimate this, we generated clusters of sequences, which were selected at random from the set of 45 sequences and calculated the resulting conservation scores. High scores for randomly selected clusters are expected to be a reflection of a high degree of redundancy within the selected set of sequences or an insufficient number of sequences in the analyzed cluster. Figure 4 shows the distribution of scores obtained for 500 randomly selected clusters of various sizes, ranging from 5 to 25 sequences. During this analysis, the weight assigned to each sequence was recalculated for each randomly compiled cluster. With an increasing number of sequences, the distribution of conservation scores narrows and becomes less noisy as the number of events per bin increases. For the smallest randomized cluster, containing only five sequences, sporadic high scores up to 0.5 are observed. With 10 sequences in a cluster, this threshold drops to 0.2 and with 15 sequences, the size of the heregulin cluster in the narrowest partition used in our analysis, the random score distribution has virtually dropped to zero at 0.15. Based on this analysis, we set the threshold for any specific conservation score to be considered a true signal to 0.2 for partition IV and III and 0.15 for

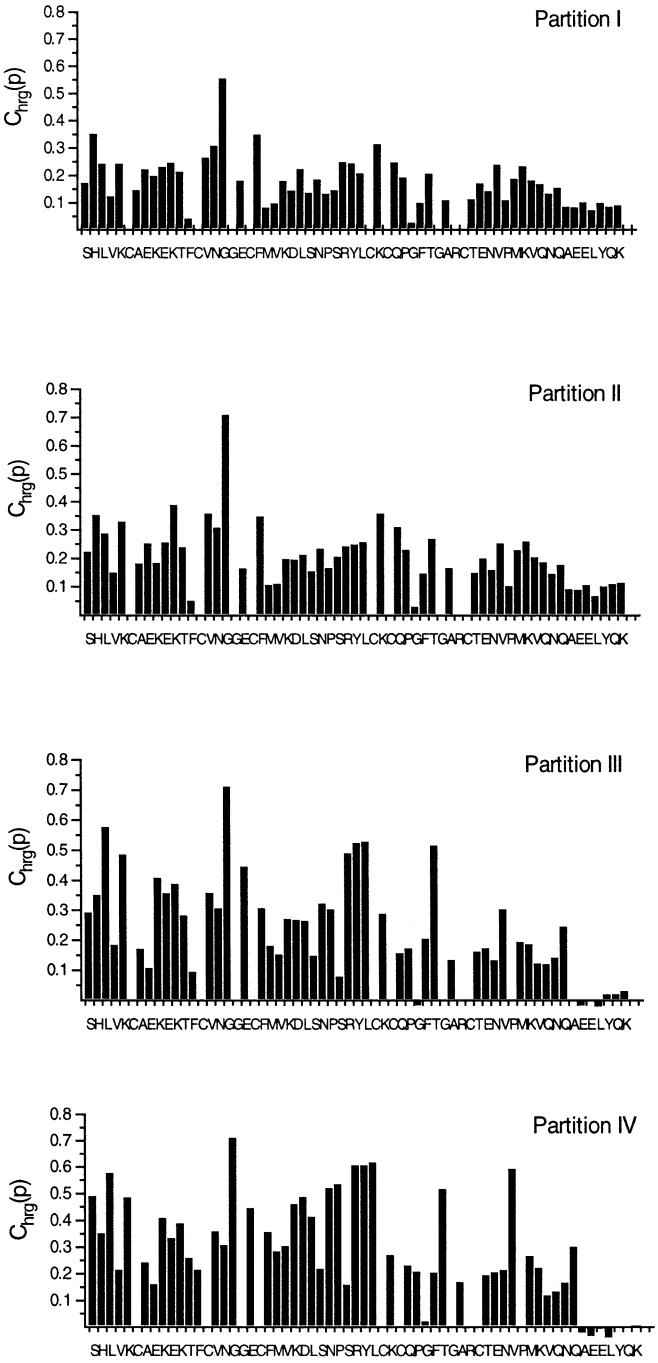

Fig. 3. Specific conservation (C_{hrg}) of the hrg cluster at partitions **I–IV**. The heregulin specific conservation scores (C_{hrg}) for each partition are presented against the sequence of heregulin alpha. Positions for which there is no equivalent in heregulin alpha have been omitted. The heregulin-specific conservation scores increase with the partition numbers, that is, as the hrg cluster under analysis decreases in size and complexity.

partition I and II. Since the value for this threshold is a reflection of the composition and size of the overall data set, it should be redetermined for each new set of sequences.

A visualization of specific conservation scores above the selected threshold is shown in Figure 5 for each of the four partitions. Based on the NMR structure of heregulin alpha (Jacobsen *et al.*, 1996), each position that scored above the noise threshold determined above has been assigned a color ranging from yellow for low scores to red for high scores. For each partition, two views of hrg are shown which are related

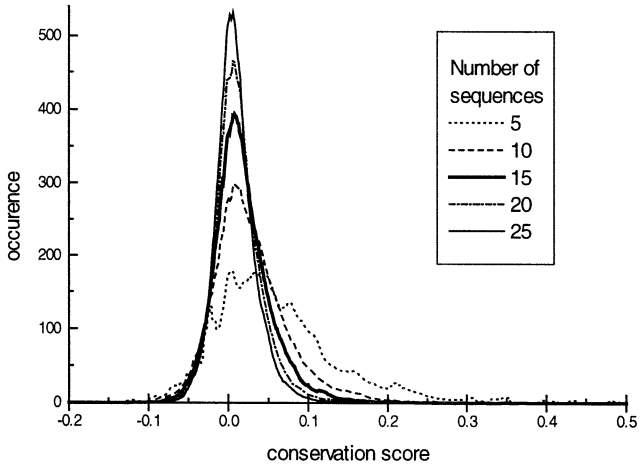

Fig. 4. Random distribution of conservation scores for different size clusters. The indicated number of sequences was randomly selected form the data set. Scores were analyzed in 0.1 size bins for 500 random clusters. Based on this analysis, the threshold for partition **III** and **IV** (16 and 15 sequences, respectively) was set to 0.2.

by a 180° rotation around the *Y*-axis. Two features stand out with respect to the overall distribution of residues with heregulin specific conservation scores.

First, at higher partition numbers high scoring positions increasingly cluster at the foot of the boot-shaped hrg molecule (Figure 5). This foot area contains the N-terminus. The C-terminus is located near the top of the boot. Figure 6 shows all scores in partition IV above the noise threshold of 0.2 and indicates the extent to which these residues are part of the N-terminal cluster. The amino acids in this cluster (black in Figure 6a, dark blue in Figure 6b) are mostly recruited from two segments of the primary sequence: the N-terminal five amino acids and the sequence surrounding the Ω-loop region of hrg. Most of the remaining residues with high scores (hatched columns in Figure 6a, light blue in Figure 6b) are related to the cluster by extending a stretch of residues on the surface of the molecule that wraps around the foot of hrg. Of the high scoring residues, only Thr41 and Val49 appear to be far removed from the cluster. Although less pronounced, this pattern is also visible in partition III. In partition I and II, high scoring positions tend to be located towards the center of the molecule.

Second, the heregulin molecule exhibits a sidedness with respect to the distribution of high conservation scores. With the exception of the foot of the molecule, the A face, depicted on the left side in Figure 5, shows a lower density of high scoring residues. This lower density is particularly pronounced in partition I and II. Most residues marked on the A face of the molecule are in fact located in between the two faces and are accessible from both the A and B face. Consequently, a central spine running along the A face is left virtually unmarked. Furthermore, changes between different partitions are more pronounced on the B face, while the general pattern observed on the A face remains relatively constant and changes only in intensity.

Discussion

Weighted evolutionary tracing

The goal of our weighted evolutionary trace analysis was to identify positions that (a) confer a distinct specificity to a subset of sequences, such as the heregulins, and (b) set this

subgroup apart from the remainder of the sequences within this family of EGF-like growth factors. Thus, the goal of our approach, which we have named weighted evolutionary tracing, differs from the original method described by Lichtarge *et al*. (1996). Evolutionary tracing, as it was initially described, 'traces' the emergence of a common sequence signature throughout a large family of proteins that are related in sequence and structure. The correlation with a three-dimensional structure is then used to identify those conserved positions which constitute a readily identifiable cluster in three-dimensional space. To this end, subgroups of sequences, as defined by the sequence dendrogram of the entire sequence family, are represented by sequence profiles that classify individual amino acids as conserved, variable or class-specific. A class-specific residue is a residue that shows conservation within all subgroups of a family but varies if the profiles for different subgroups are compared. If a position is not conserved within one ore more of the subgroups, it is considered variable. Evolutionary tracing in this form would analyze the family of EGF-like growth factors for the emergence of a cluster of conserved positions, common to this entire family of proteins

The underlying assumption to our weighted evolutionary trace approach is that positions which are conserved within the family of heregulins but are variable in the family of EGF-like growth factors as a whole are likely to contribute to the specificity of heregulins. Such group-specific conservation could manifest itself in different ways, thereby going beyond the definition of a class-specific residue used in the original version of evolutionary tracing. A position could be highly conserved within a subgroup but variable outside it. Alternatively, a different conserved residue may be found in all or a portion of the remaining sequences. Both scenarios could involve drastic or conservative variations. In order to incorporate and quantify the above scenarios in our analysis, we chose to calculate the degree of variation in each position based on the amino acid substitution values represented in the Gonet matrix.

The amount of information yielded by this analysis depends directly on the number of sequences available for analysis. A large number of sequences (>100) are available for EGF-like growth hormones. However, excluding sequences with identical EGF-like domains, the variations in the remaining sequences are minor, especially in the case of heregulins. However, while contributing little to the overall sequence diversity, rare mutations in otherwise conserved sequences represent important information. By weighting sequences by their level of sequence diversity, we intended to incorporate this information and prevent the over-representation of similar sequences. The result of this weighting scheme is evident in the data obtained for the different partitions and heregulin clusters. The heregulin clusters in partition I and II differ by only one sequence: mouse neuregulin 3. The high weight assigned to this sequence compared to the more similar heregulins in the center of the cluster results in a notable change in the pattern of high scoring residues (Figure 5).

Weighted evolutionary trace analysis of heregulin

For our analysis of heregulin, we selected 45 sequences from the family of EGF-like growth factors. These sequences were

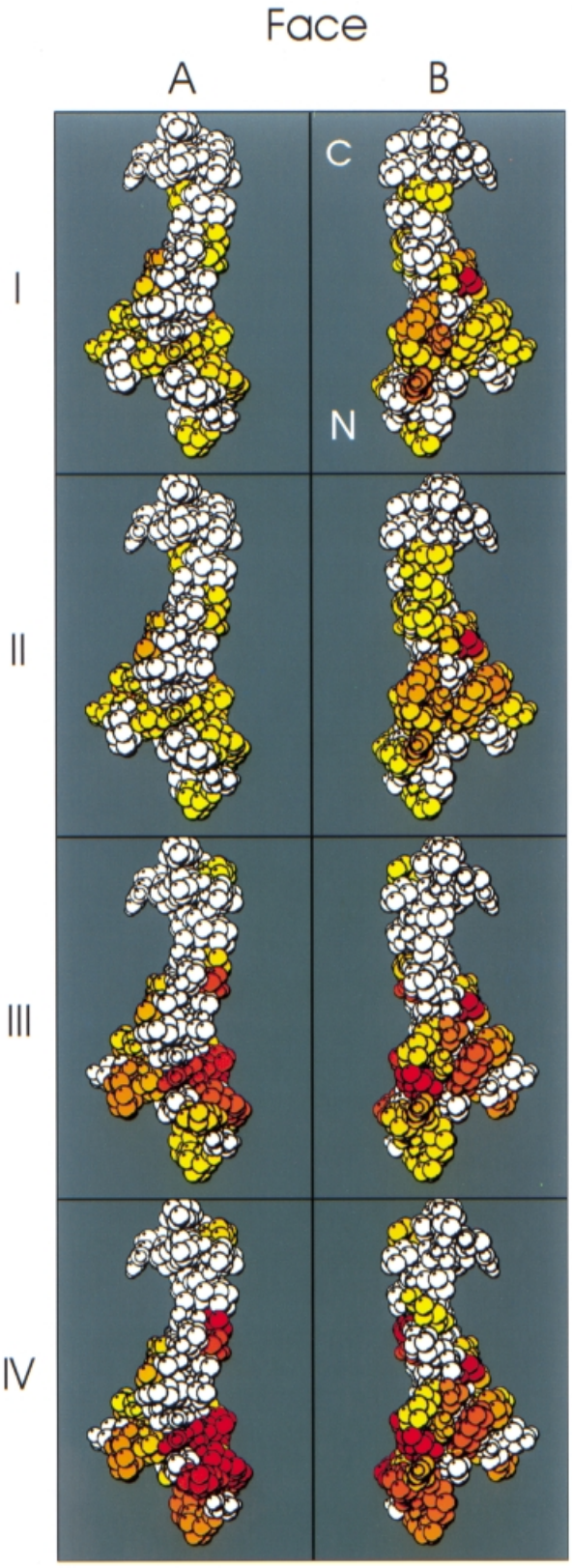

Fig. 5. Distribution of heregulin-specific conservation scores. The C_{hrg} scores are superimposed on the NMR structure of heregulin alpha. Scores above the noise threshold are shown using a color gradient from yellow (equal to noise threshold) to red (maximum value in this partition). The two opposite faces (A and B) of heregulin alpha are shown for partition **I–IV** (as indicated to the left). Both faces are related by a 180° rotation around the *Y*-axis. The foot of the boot-shaped heregulin domain points down in both representations. In partition **I**, the C- and N-termini are indicated as 'C' and 'N′ respectively. As the partition number increases, positions with high C_{hrg} scores are found with greater abundance at the foot of hrg. An uneven distribution of high scores between the A and B face is especially pronounced at partition **I** and **II**.

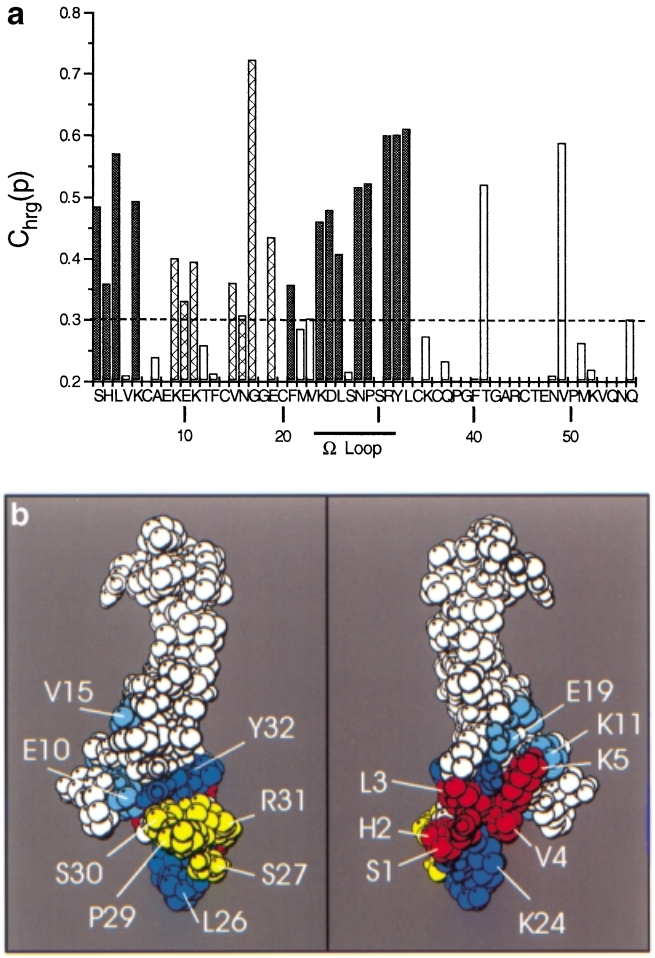

Fig. 6. (**A**) Distribution of positions with high heregulin specific conservation scores (C_{hrg}). High scoring positions in partition **IV** cluster at the foot of heregulin and are located in the N-terminal and Ω-loop segments. Except for T41 and V49, all positions with high relative conservation scores (>0.3) can be viewed as part of an N-terminal cluster [black in (A), blue in (B)] or as an extension of it [hatched in (A) and light blue in (B)]. (**B**) Visualization of positions marked in (A). The core of the N-terminal cluster consists of positions within the N-terminus and Ω-loop region of heregulin. The N-terminal ('SHLVK', red) and Ω-loop ('SNPSR', yellow) core motif is found on opposite faces of the molecule.

obtained by searching the Genbank database with the sequence of human EGF and heregulin α. For those non-redundant sequences with *P*-scores below 10^{-5}, two additional criteria were used to reduce further the number of sequences used for evolutionary tracing. A wide variety of EGF-like domains can be found in proteins without growth factor function. We limited the sequences in this analysis to those proteins known to have growth or differentiation factor activities. This selection resulted in several sequences, which are identical within the 63 amino acid window, determined by the EGF-like domain of heregulin alpha. Those sequences were kept to demonstrate the level of conservation within a class of ligands. The resulting redundancy as well as the small variations among almost identical sequences were taken into account by the weighting scheme. Biologically active peptides represent an important group of sequences, which are missing from our analysis. The published biochemical analysis of these peptides is often limited to a specific receptor in the HER/EGFR family or a single biochemical property, such as binding. It is therefore difficult to evaluate the contribution of these sequences to the detection of specificity conferring residues. An evolutionary trace analysis which included a set of peptides, characterized by Barbacci *et al.* (1995), produced the same results as were obtained with exclusively biological ligands (data not shown). However, such peptides often represent the product of ligand design rather than natural selection and their addition can potentially limit the interpretation of the results of this analysis.

The effect obtained by the omission of sequences without confirmed growth factor activity is difficult to determine. EGF-like domains are found in a large number of proteins, most of which are not implicated as ligands for growth factor receptors. The inclusion of highly divergent sequences into the evolutionary tracing analysis could potentially help in the identification of functionally important positions among the known growth factors but may violate the underlying assumption of close structural similarity, implicit in our analysis. The model will also be influenced by the inclusion of new EGF-like growth factors.

The underlying assumption of the partitioning scheme is that the separation of clusters at different partitions within the dendrogram gives an approximate representation of the evolution of growth factors with different specificity. For the selected sequences, this assumption is, for the most part, confirmed by the dendrogram in Figure 1. However, the resulting separation into an EGF cluster and a heregulin cluster has to be interpreted with caution. For example, betacellulin has been reported to activate HER2/HER3 heterodimers (Alimandi *et al.*, 1997) and the activation of MCF7 cells by EGF can involve HER2 (Karunagaran *et al.*, 1996). However, despite cross-activation, this cluster assignment does reflect the dominance of some receptor interaction over others and most ligands in the heregulin cluster share HER2 involvement as the common denominator for maximum activation (Graus-Porta *et al.*, 1997). Thus, the resulting partition in Figure 1 should not be interpreted as reflecting absolute specificity but rather binding and activation preferences.

The overall variability scores shown in Figure 2 indicate a conserved core within the first 50 amino acids, both for the heregulin cluster and all combined sequences. This result is consistent with findings that truncations of the C-terminal segment of hrg are tolerated without apparent loss of activity (Barbacci *et al.*, 1995). The equivalent C-terminal segment is absent in mature TGF alpha and some forms of EGF and most artificial peptide agonists represent C-terminal truncations of the respective wild type growth factor.

The values of C_{hrg} (Equation 3) reveal high scoring residues throughout the sequence of heregulin, but superimposition on the three-dimensional structure of heregulin alpha reveals clustering near the foot of the boot-shaped hrg domain. This is especially the case at partition III and IV. The residues are recruited to this cluster primarily from two regions in the primary sequence: the N-terminus and the Ω-loop (Figure 6). These two portions of the molecule also stand out in the NMR structure as being the most flexible portions within the well ordered first 50 amino acids of hrg (Jacobsen *et al.*, 1996). Although the N-terminus has been recognized in the past as an important element in conferring receptor specificity (Barbacci *et al.*, 1995), the Ω-loop has not. In fact, the deletion of amino acids within the Ω-loop does not seem to reduce binding to HER3 (Barbacci *et al.*, 1995).

The sequence and biological properties of NTAKs and neuregulin 2 and 3 are interesting in this respect. While all three are clearly HER ligands, their sequences show significant

differences to the sequences of the core heregulin cluster at partition IV, primarily in the Ω-loop. With the single exception of human heregulin γ, deletions in the Ω-loop are a hallmark of ligands in the EGF cluster. A comparison of the biological activity of NTAKs (Higashiyama *et al*., 1997) and neuregulin 2 (Carraway *et al*., 1997) and 3 (Zhang *et al*., 1997) with typical heregulins indicates a difference in receptor specificity. All of these 'non-typical' heregulins seem to feature a stronger dependence on HER3 and especially HER4 compared with HER2. Thus the emergence of the cluster at the foot of hrg in partition III and IV, both of which exclude the above ligands, may be a reflection of these differences in receptor preferences.

An interesting feature of the two sequence segments in the N-terminus and Ω-loop is the presence of a weak sequence similarity. The N-terminal five amino acids (SHLVK), confirmed to be crucial for high affinity binding to HER3, have a distant counterpart in the central five amino acids (SNPSR) of the Ω-loop segment. Both segments are located on opposite faces of the heregulin molecule (Figure 6b). The N-terminal (SHLVK) motif runs along face B of heregulin. The presence of the central proline in the Ω-loop motif (SNPSR) makes this segment form a non-extended patch on the opposing A face. The placement of the Ω-loop motif on the opposite face of the N-terminal motif, known to be important for HER3 binding, may indicate a role as a negative regulatory element. The function of this sequence element in the Ω-loop, which is not required for binding to HER3 or HER4, may in fact be to interfere with interactions of the A face of hrg to receptors that bind strongly to the B face. As a result, it would disfavor binding of hrg to homodimers of HER3, HER4 or EGFR and instead favor the formation of HER2 containing heterodimers. Support for this hypothesis comes from the alanine scanning results of heregulin β. Overall, five out of eight positions, at which a mutation to alanine increased affinity to either HER3 or HER4, fell into the Ω-loop region. Specifically, the mutation of the central proline (position 29) in the Ω-loop motif to alanine was the single most potent mutation to increase binding to HER4. The strongest gain in binding to HER3 was obtained by the replacement of serine 27 at the beginning of the Ω-loop motif to alanine (Jones *et al*., 1998). While proline 29 shows a strong heregulin-specific conservation score, serine 27 scores only marginally above the noise level. A possible explanation for this difference could be that the unique properties of proline are required at position 29, while the requirements for position 27 are broader. While position 27 contains a positively or negatively charged residue in all ligands of the EGF cluster, it varies between threonine, proline and serine in the heregulin cluster (data not shown). Such a weakly defined pattern would not be picked up based on the current scoring scheme, which relies on a Gonnet substitution matrix.

An overall comparison of the weighted evolutionary trace analysis with the alanine scanning results (Jones *et al*., 1998) shows a strong correlation for the N-terminus of heregulin, up to position 20 (data not shown). When comparing the results of the alanine scan of heregulin β (Jones *et al*., 1998) with the weighted evolutionary trace analysis, one has to keep in mind that the methods use an entirely different readout. Alanine scanning identifies all positions important to ligand function but does not necessarily distinguish between residues of global importance to the entire family of EGF-like ligands and residues that confer specificity to the hrg-HER interaction. Consequently, the majority of amino acids identified by alanine scanning were classified as being important to EGF-like ligands in general. Different results would be expected from the evolutionary trace analysis in regions of the molecule whose primary function is not to increase binding affinity but to modulate specificity, such as the central Ω-loop region, which stands out in the alanine scanning analysis by its propensity for mutants with increased binding.

Conclusion

A weighted evolutionary trace analysis of the hergulin family compared with the entire family of EGF-like growth factors suggests that the hrg domain contains two functionally distinct surfaces which we termed face A and face B. Since this analysis is designed to uncover group-specific conservations within a family of proteins, the two functionally distinct interfaces may reflect differences in specificity between hrg ligands and the EGF-like ligands as a whole. Besides the preference for a different subset of receptors (HER2, 3 and 4 versus EGFR), hrg also shows a strong preference for the interaction with receptor heterodimers versus the homodimeric interaction seen between EGF and EGFR.

Our analysis further suggests a clustering of crucial residues at the foot of the hrg domain. The residues that constitute this cluster are located in the primary sequence at the N-terminus and Ω-loop and show a weak sequence similarity. The N-terminal residues known to be important for the high affinity interaction with HER3 are located on the B face. Overall, the B face shows a greater density of high-scoring positions and is likely to represent the high-affinity interface with HER3. The corresponding positions in the Ω-loop (face A) are more likely to act as a negative regulatory element which might control preference of hrg for receptor heterodimers over homodimers and may represent the low-affinity HER2 interface. This hypothesis is supported by the fact that mutations to alanine in this region have been reported to increase binding to HER3 or HER4 (Jones *et al*., 1998). Weighted evolutionary tracing, as it is presented in this work, provides a simple method to correlate the divergence in sequence space among a family of related proteins to divergence in function.

Acknowledgements

This work was supported by the National Institutes of Health [Grant GM31299 (D.E.) and NRSA 1F32 AI09619 (R.L.)]. We thank Drs Mark Pegram, Danny Rice and Lukasz Salwinski for helpful discussions and the NIH and the Lita Annenberg Hazen Foundation for support.

References

Alimandi,M., Wang,L.M., Bottaro,D., Lee,C.C., Kuo,A., Frankel,M., Fedi,P., Tang,C., Lippman,M. and Pierce,J.H. (1997) *EMBO J*., **16**, 5608–5617.

Barbacci,E.G., Guarino,B.C., Stroh,J.G., Singleton,D.H., Rosnack,K.J., Moyer,J.D. and Andrews,G.C. (1995) *J. Biol. Chem*., **270**, 9585–9589.

Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) *Protein Engng*, **7**, 1323–1332.

Carraway,K.L.R., Weber,J.L., Unger,M.J., Ledesma,J., Yu.N., Gassmann,M. and Lai,C. (1997) *Nature*, **387**, 512–516.

Casari,G., Sander,C. and Valencia,A. (1995) *Nature Struct. Biol*., **2**, 171–178.

Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res*., **12**, 387–395.

Earp,H.S., Dawson,T.L., Li,X. and Yu,H. (1995) *Breast Cancer Res. Treat*., **35**, 115–132.

Gassmann,M., Casagranda,F., Orioli,D., Simon,H., Lai,C., Klein,R. and Lemke,G. (1995) *Nature*, **378**, 390–394.

Graus-Porta,D., Beerli,R.R., Daly,J.M. and Hynes,N.E. (1997) *EMBO J*., **16**, 1647–1655.

Guy,P.M., Platko,J.V., Cantley,L.C., Cerione,R.A. and Carraway,K.L.R. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 8132–8136.

Han,X., Kasahara,N. and Kan,Y.W. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 9747–9751.
Higashiyama,S. *et al*. (1997) *J. Biochem*. (Tokyo), **122**, 675–680.
Holmes,W.E. *et al*. (1992) *Science*, **256**, 1205–1210.
Jacobsen,N.E., Abadi,N., Sliwkowski,M.X., Reilly,D., Skelton,N.J. and Fairbrother,W.J. (1996) *Biochemistry*, **35**, 3402–3417.
Jones,J.T., Ballinger,M.D., Pisacane,P.I., Lofgren,J.A., Fitzpatrick,V.D., Fairbrother,W.J., Wells,J.A. and Sliwkowski,M.X. (1998) *J. Biol. Chem*., **273**, 11667–11674.
Karunagaran,D., Tzahar,E., Beerli,R.R., Chen,X., Graus-Porta,D., Ratzkin,B.J., Seger,R., Hynes,N.E. and Yarden,Y. (1996) *EMBO J*., **15**, 254–264.
Kihara,A. and Pastan,I. (1995) *Cancer Res*., **55**, 71–77.
Kohda,D. and Inagaki,F. (1992) *Biochemistry*, **31**, 11928–11939.
Landgraf,R., Pegram,M.D., Slamon,D.J. and Eisenberg,D.S. (1998) *Biochemistry*, **37**, 3220–3228.
Lee,K.F., Simon,H., Chen,H., Bates,B., Hung,M.C. and Hauser,C. (1995) *Nature*, **378**, 394–398.
Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) *J. Mol. Biol*., **257**, 342–358.
Marchionni,M.A. *et al*. (1993) *Nature*, **362**, 312–318.
Park,J.W. *et al*. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 1327–1331.
Pegram,M.D., Finn,R.S., Arzoo,K., Beryt,M., Pietras,R.J. and Slamon,D.J. (1997) *Oncogene*, **15**, 537–547.
Plowman,G.D., Culouscou,J.M., Whitney,G.S., Green,J.M., Carlton,G.W., Foy,L., Neubauer,M.G. and Shoyab,M. (1993a) *Proc. Natl Acad. Sci. USA*, **90**, 1746–1750.
Plowman,G.D., Green,J.M., Culouscou,J.M., Carlton,G.W., Rothwell,V.M. and Buckley,S. (1993b) *Nature*, **366**, 473–475.
Sibbald,P.R. and Argos,P. (1990) *J. Mol. Biol*., **216**, 813–818.
Siegall,C.B. *et al*. (1995) *J. Biol. Chem*., **270**, 7625–7630.
Slamon,D.J. *et al*. (1989) *Science*, **244**, 707–712.
Sliwkowski,M.X. *et al*. (1994) *J. Biol. Chem*., **269**, 14661–14665.
Tzahar,E. *et al*. (1994) *J. Biol. Chem*., **269**, 25226–25233.
Tzahar,E. *et al*. (1997) *EMBO J*., **16**, 4938–4950.
Zhang,D. *et al*. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 9562–9567.

Received February 4, 1999; revised July 19, 1999; accepted April 8, 1999