CABIOS

Vol. 12 no. 3 1996
Pages 197–203

CONRAD: a method for identification of variable and conserved regions within proteins by scale-space filtering

Günter Herrmann[1], Annemarie Schön[2], Ruth Brack-Werner[2] and Thomas Werner[3]*

Abstract

Advanced sequencing techniques allow rapid deduction of individual amino acid sequences of highly related proteins. Due to their quasi-species nature, viral genomes (e.g. HIV-1) represent one of the most common sources of related proteins. Another example of related proteins are immunoglobulins. Local differences in amino acid conservation are useful indicators of potential domain structures and immunological or functional epitopes prior to structural analysis of proteins. Although variability indices can be calculated by several methods, delineation of boundaries between sequence stretches with similar variability indices is left to the user. We use algorithmic scale-space filtering for delineation of conserved and variable sequence stretches within a protein which is performed on an algorithmic basis avoiding arbitrary assignments. Our method correctly identified variable regions for the human immunoglobulin λ-chain V-regions (subgroup I). Prediction of the variable regions of the HIV-1 gp120 env protein was in agreement with empirical derived definitions. These examples indicate that our method is useful for the regional assignment of protein variability solely on the basis of amino acid sequences.

Introduction

Large scale DNA sequencing reveals open reading frames of proteins much faster than analysis of the corresponding protein sequences can proceed. Especially viral protein sequences deduced from PCR-sequencing of viral genomes represent rapidly growing families of highly related protein sequences. The information represented by local differences in amino acid conservation can be used to deduce potential domain structures as well as to identify candidates for immunological or functional epitopes. Examples for such approaches are the definition of the immunoglobulin hypervariable regions (Kabat and Wu, 1970; Kabat *et al.*, 1991) or the domain definition of the HIV-1 gp120 protein by Modrow *et al.* (1987). In the meantime some programs are available to aid scientists in these predictions (e.g. 'PlotSimilarity', in Genetics Computer Group, 1994; or 'VIR', Almagro *et al.*, 1994).

However, though positional consensus scores are calculated by those methods, prediction of region boundaries is left to the user. This can be a major obstacle for regions with only modest variability since prediction of domains is restricted to those regions which are clearly identifiable by visual inspection of variability indices.

Here we present an algorithmic method that is capable of delineating region boundaries systematically by scale-space filtering. We show our method to correctly identify hypervariable regions for the example of human immunoglobulin λ-chain V-regions (subgroup I) and to predict the variable regions of the HIV-1 gp120 in agreement with empirical definitions. The method is not restricted to multiple alignment analysis. Similar signals for individual sequences like hydrophobicity or probability profiles from secondary structure prediction methods can also be analysed.

Algorithm

Our method CONRAD (CONserved RAnge Detection) is designed to allow complete dissection of the input signal (e.g. a consensus score) into 'conserved' and 'variable' regions by systematic filtering solely based on the input signal. Our algorithm employs a consensus score obtained from the multiple alignment as 'consensus signal' for the filtering process. Currently, we use the similarity values assigned by the GCG program PlotSimilarity. However, any method yielding consensus scores can provide the consensus signal. A calculated threshold score (usually the average of a consensus score) will be used to separate constant and variable regions. Insignificant fluctuations of the original signal around the threshold obscuring the region boundaries are removed by filtering which represents smoothing of the signal.

Basic principles

Our algorithm is based on the method of scale-space filtering originally published by Witkin (1983). The basic idea is that prominent features in a signal or image are visible already at low resolutions and keep their

[1]*Institut für Medizinische Informatik und Systemforschung,* [2]*Institut für Molekulare Virologie and* [3]*Institut für Säugetiergenetik, GSF-Forschungszentrum für Umwelt und Gesundheit GmbH, Ingolstädter Landstraße 1, D-85758 Oberschleißheim, Germany*

**To whom correspondence should be addressed. E-mail: werner@gsf.de*

appearance over a broad range of resolutions. Scale-space filtering models an intuitive approach: First the signal is smoothed repeatedly with low-pass filters (weighted average filters) of increasing strength, until all details disappear. Then the whole range of resolutions (scales) is analysed for features (minima or maxima) that remain stable over a large range of scales. Boundaries of these features are back traced to the original signal. Thus, scale-space filtering is an algorithmic solution for phenomena which can be considered as diffusion problems. For excellent overviews of this methodology see Witkin (1983), Lindeberg (1990, 1993), Yuille and Poggio (1986), and references cited therein.

Scale-space filtering

'Scale space' is the embedding of the original signal $f(x)$ into a family of derived signals $L(x;t)$ all of which were generated using a family of low-pass filters $\{T(x;t)\}$:

$$L(x;t) = f(x-n)T(-n;t) + f(x-n+1)T(-n+1;t) + \ldots + f(x+n)T(n;t)$$

The scale parameter t is the filter standard width (or any monotonic function). The filter family $T(x;t)$ must fulfill strict restrictions: (i) no additional extrema (not present in the unfiltered signal) may appear during the filter process, and (ii) for any t, $L(x;t)$ depends only on $f(x)$ and t, independent of any intermediate filter stages.

This has a clear consequence for the form of the filter family. The $L(x;t)$ must be solutions of a diffusion equation (Babaud *et al.*,1986; Yuille and Poggio, 1986; Liu and Rangayyan, 1991). Despite of the continuous nature of scale, it is sufficient for our purpose to consider only discrete

```
num.a    1                                                        50
l1hung   qsvltqppsv saapgqevti scsgsssnig dnf.vswyqq lpgtapklli
l1hubl   qsvltqppsv saapgqkvti scsgsssnig ndy.vswyqq vpgtapklli
l1hunw   qsvltqppsv saapgqkvti scsggstnig nny.vswhqh lpgtapklli
l1huep   qsvltqppsl saapgqrvsi scsgsssnig kny.vqwyqq lpgtapklli
l1huwa   qsvltqppsa sgtpgqrvti scfgsssnig ryy.vywyqq lpgttpklli
l1humm   qsvltqppsa sgtpggrvti scsgsssnvg snzpaywyqq lpgtapklli
l1huha   qsvltqppsv sgtpgqrvti scsggssngt gnnyvywyqq lpgtapklli
l1huvo   qsvltqppsa sgtpgqrvti scsggnfdig rn.svnwyqv hpgtaprlli
l1hunm   qsvltqppsv sgapgqrvti scsgsssnig agnhvkwyqq lpgtapklli
num.b                          23              35
                                  CDRI

num.a    51                                                       100
l1hung   ydnnkrpsgi pdrfsgsksg tsatlgitgl qtgdeadyyc gtwdsslsvg
l1hubl   ydnnkrpsgi pdrfsgsksg tsatlgitgl qtgdeadyyc gtwnnslsgw
l1hunw   yednkrpsgi pdrisasksg tsatlgitgl rtgdeadyyc atwdsslnav
l1huep   fnnnkrpsgi pdrfsgsksg tsatlgitgl qtgdeaiyyc gtwdnrrs..
l1huwa   ykdnqrpsgv pdrfsgsksg tsaslaisgl rsedeadyyc aawddsl..w
l1humm   ynynqrpsgv pdrfsasrsg tsaslaisgl qsedeadyyc aawddsldgy
l1huha   yrddkrpsgv pdrfsgsksg tsaslaisgl rsedeahyhc aawdyrlsav
l1huvo   yssdqrssgv pdrfsgsksg tsaslaisgl qseneadyfc atwddsldgp
l1hunm   fhnn....... .arfsvsksg ssatlaitgl qaedeadyyc qsydrsl..r
num.b    49     57                                  88
           CDRII                                        CDRIII

num.a    101       112
l1hung   mfgggtrvtv lg
l1hubl   vfgggtkltv lg
l1hunw   vfgggtkvtv lg
l1huep   vfgggtnvtv vg
l1huwa   vfgggttltv ls
l1humm   vfgtgtkvtv lr
l1huha   vfgggtqltv lr
l1huvo   vfgggtkvtv lg
l1hunm   vfgggtkltv lr
num.b    98
```

Fig. 1. Multiple alignment of V-regions (subgroup I) of human immunoglobin λ-chain gene. Multiple alignment of 9 human immunoglobulin λ-chain V-regions (subgroup I; all entries from PIR-database, release 46.0, Barker *et al.* 1994) was carried out by GCG PileUp. Sequences are labeled with their PIR-database identifiers. CDR regions are boxed; *num a*: numbering corresponds to PileUp (GCG 1994); *num b*: numbering corresponds to Kabat *et al.* (1991).

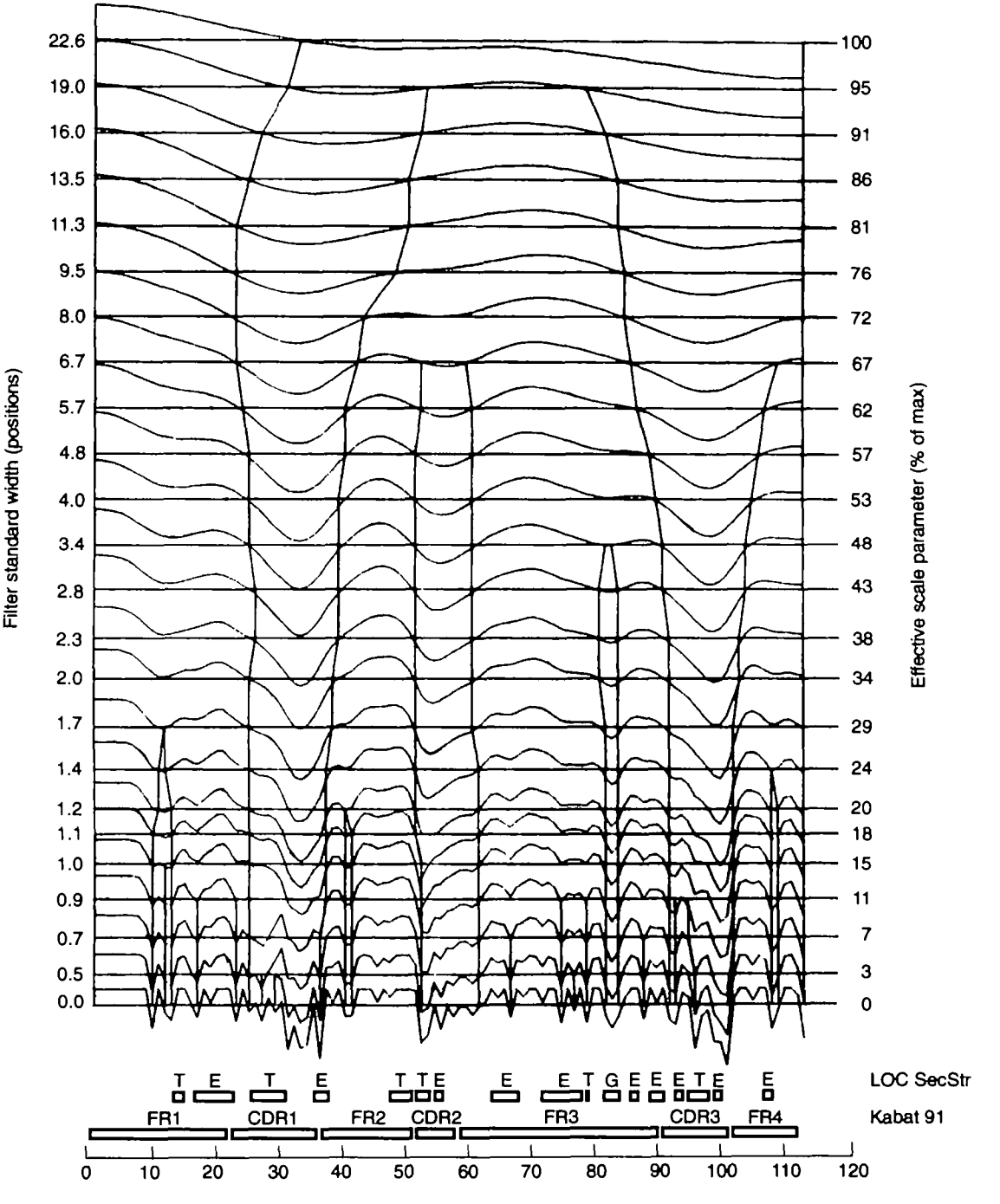

Fig. 2. Space-scale filtering of V-regions (subgroup I) of human immunoglobin λ-chain gene (see Fig. 1). The numbers at the left vertical axis represent the filter standard width (in positions) for the respective filter level. The numbers at the right axis represent the corresponding effective scale (in normalized units). The distances of the plotted axes at each filter level from the zero level axis are linear proportional to the effective scale. The unfiltered variability plot is shown at the zero-level, below this the amino acid positions of the consensus are shown. Two region classifications from the literature are indicated as horizontal bars for comparison (not present in program output). CDR regions are taken from Kabat *et al.* (1991) and marked as 'Kabat 91'. Secondary structure interpretations of Bence Jones Protein LOC (Schiffer *et al.*, 1991) with program DSSP (Kabsch and Sander, 1983) are marked 'LOC sec.str.' (E: beta-strand, T: H-bonded turn, G: three-helix). The LOC sequence is not part of the input sequences.

levels of t with preselected steps of sufficiently fine resolution, in accordance with Liu and Rangayyan (1991). In this case, we can use a family of generalized binominal filters.

In our program, we start with a 3-tap filter of standard width 0.5 as basic filter. The filter family converges quickly to Gaussian filter kernels which are used as approximations for large t. The signal is mirrored at both ends for filtering, as suggested by the ideas behind the discrete cosine transformation (see Blinn, 1993).

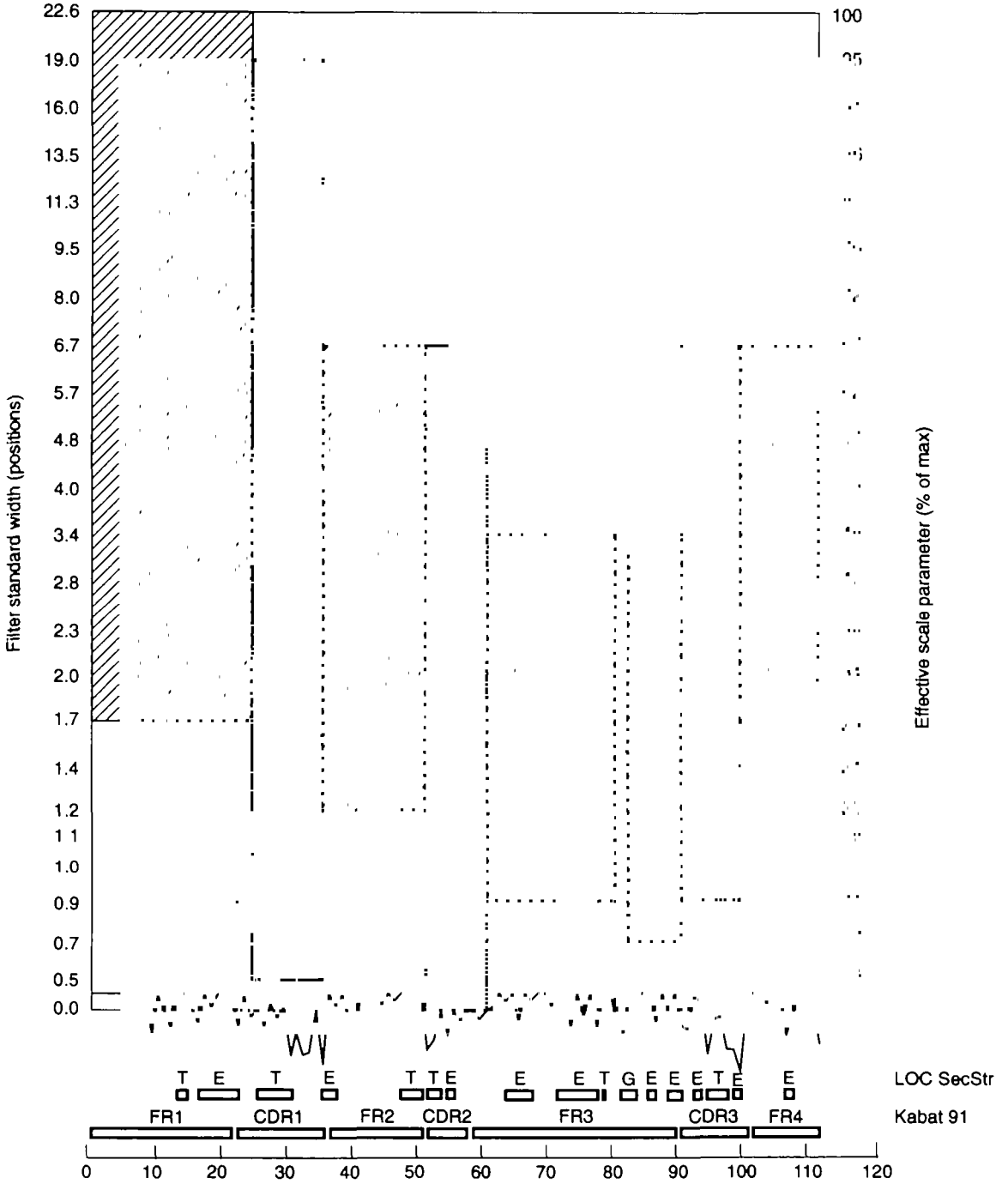

Fig. 3. Region plot of the space-scale filtering from Fig 2. Scales left, right, and below are the same as in Fig. 2. Classification into regions was based on the fingerprint from Fig. 2. The horizontal extensions of the rectangles refer to region start and end positions on the level of unfiltered signal. The height represents the 'life-time' in scale space. C-regions determined by the algorithm as most stable are hatched, V-regions are cross-hatched. Blank boxes represent less stable regions with shorter 'life-times' than the most stable regions.

Definition of regions

For the scope of this paper, a region shall be defined as a sequence stretch between consecutive threshold crossings of opposite sign, at any filter level. Each region exists during a certain 'lifetime' in scale space, and splits at its lower level of scale into three sub-regions, thus forming a ternary tree. Figure 1 shows initial the multiple sequence alignment on which our analysis was based. At each filter level, the threshold crossings must be determined and aligned to the threshold crossings of the next lower filter level. The resulting contours of connected threshold crossings in the signal-scale plane form the so-called 'fingerprint'. An example for this is shown in Figure 2. Merging of three neighboring regions can be observed in the fingerprint each time two adjacent contours join to an arc closed at the top (see Fig. 2). Begin and end of a region in signal space are referenced to positions on the unfiltered signal by tracing the fingerprint contours down to filter level 0. All possible regions are plotted as rectangles in Figure 3.

Since we restrict to discrete filter levels, the aligning of the threshold crossings at adjacent filter levels has to be carried out with some precautions, using linear prediction from lower filter levels and suitable level-dependent tolerances for fit. We obtain a consistent tree of regions in practically all cases if the distances between filter levels are not too large.

Effective scale

Lindeberg (1993) introduced the term 'effective scale' as a monotonic function $s(t)$ of the ordinary scale parameter t in order to provide a measure that is suitable for direct comparison of stabilities (life-times). His approach is based on the assumption that for a random noise signal the relative decrease of local extrema should be independent of effective scale. He also gives a formula for $s(t)$ by estimating the filter dependent density of extrema on random noise. Lindeberg's formula is valid for any scale down to the unfiltered signal. For large scales, $s(t)$ converges to the logarithm, with evidence that zooming of a waveform in signal space results in a shift in effective scale space with differences in scale preserved. For details, see Lindeberg (1993).

In Figures 2 and 3, the vertical axis is drawn linear in effective scale (upper limit set to 100%). Numbers at the left side give the standard filter width t of each filter level. The first 13 filter levels are realized with members of the family of generalized binominal filters described above with filter orders chosen to obtain fairly equal steps in effective scale. Higher filter levels are approximated with Gaussian filters and the logarithm as effective scale.

Definition of a 'most stable region'

A region of 'maximum local stability' is defined as a region that contains no subregions with a longer life-time, and itself is no subregion of another region with a longer life-time (Witkin, 1983). Our method constructs a complete dissection of the whole signal into regions of maximum local stability using a two-step strategy based on the above-mentioned ternary tree: In a first step, recursively from bottom up, all sub-branches containing only less stable nodes are deleted. In a second step, from top down, all nodes marked from step one as containing sub-branches with more stable nodes are deleted.

Environment

CONRAD was written in C and has been tested on a DEC Alpha Station under OSF/1 V3.2 as well as on a PC-486 under FreeBSD 2.0 with gcc. Graphical output is produced either as a PostScript or a HPGL file. A summary of the results is written to an ASCII output file.

Results and Discussion

The purpose of our method is a complete dissection of the input signal (e.g. a consensus sequence) into 'conserved' and 'variable' regions by systematic filtering. Our method is free of user-defined parameters and does not require any information except an input signal which can be derived from a set of related amino acid sequences (multiple alignment) or by analysis of a single sequence (e.g. hydrophobicity or other physico-chemical property plots, secondary structure probability values, etc.). The algorithm is independent of the amino acid sequences and is not influenced by compositional biases of the protein(s). Thus it can be used to compare variability profiles between different protein families.

The input for our program is an array of any kind of variability values for each position of the consensus-sequence. Here, we used a multiple alignment of sequences (GCG PileUp) from which an array of variability indices is deduced (GCG PlotSimilarity, window = 1). However the program can accept any kind of variability indices provided values are assigned to each consensus position.

We applied our method to the well known human immunoglobulin λ-chain V-regions (subgroup I) in order to define the complement determining regions (CDR). Figure 1 shows the multiple alignment of nine human immunoglobulin λ-chain V-regions (subgroup I) created by the GCG program PileUp. As shown in Figures 2 and 3 and inTable I our program correctly identified all three CDR regions which correspond to 'turn-regions' (T) in the secondary structure. There is an additional 'turn-region' in position 79 (numbering according to alignment), followed

Table I. Application of CONRAD region determination to human immunoglobulin λ-chain V-regions (subgroup I)

Regions CONRAD[a]	Regions Kabat[b]	Loops Chotia[c]	Secondary structure of LOC protein[d]
C 1–24	FR1. 1–22		T: 14,15, E: 17–23
V: 25–36	CDR1: 23–36	25–34	T: 26–31
C: 37–51	FR2: 37–51		E: 36–38, E: 48–51
V: 52–61	CDR2: 52–58	52–54	T: 52–54, E: 55–56
C: 62–80	FR3 59–90		E: 64–68, E: 72–78, T: 79
V: 81–83			G: 82–84
C: 84–90			E: 86–87, E: 89–91
V: 91–100	CDR3: 91–101	93–100	E: 93–94, T: 95–96, E. 99–100
C: 101–112	FR4. 102–112		E: 107–108

[a]Numbering corresponds to PileUp multiple sequence analysis program in GCG (1994).
[b]Kabat *et al.* 1991, FR = framework (conserved regions), CDR = complement determining region (hypervariable regions)
[c]Chothia *et al.*, 1986.
[d]Schiffer *et al.*, 1991, For explanation of other abbreviations see legends to Figures 1 and 2

by a three-helix structure (G) in position 82–84 which corresponds to the fourth variable region detected by our program. Though signal mirroring introduces artificial ends this did not seem to interfere with the predictive capabilities of our method in both examples and thus seems acceptable.

As shown in Table II, our method also predicted correctly the variable and constant regions of HIV-1 gp120 (seven sequences) as compared to the empirical results of Modrow *et al.* (1987) and the regions defined by Myers *et al.* (1994) (V1 and V2-loop are not separated). The functional significance of the data was confirmed by studies of the disulfide binding pattern (Leonard *et al.*, 1990) and with monoclonal antibodies (Moore *et al.*, 1994).

Table II. Application of CONRAD region determination to HIV-1 gp120 sequences

Regions CONRAD[a]	Regions Modrow 1987[b]	Regions Myers 1994[c]
C: 42–138	C1: 42–138	HD: 45–55 (gp120/start 42)
V: 139–227	V1. 139–175	V1-loop: 138–172
		HD: 132–138
	V2: 176–227	V2-loop: 174–224
C. 228–328	C2: 228–328	HD: 230–242
		HD: 270–286
V: 329–358	V3–1: 329–356	V3-loop: 326–360
C: 359–375	V3–2: 357–422	
V: 376, 377		
C: 378–383		
V: 384, 385		
C: 386–391		
V: 392, 393		
C: 394–415		
V: 416–445	V4: 423–445	V4-loop: 416–447
C: 446–489	C3: 446–489	HD: 479–484
V: 490–496	V5: 491–501	V5: 490–495
C: 497–542	C4. 502–542	HD: 503–519
		HD: 526–537 (gp 120/end 542)

[a]Numbering corresponds to PileUp multiple sequence analysis program in GCG (1994).
[b]Empirical evaluation of regions by Modrow *et al.* (1987).
[c]HD = regions of high density of information defined by Myers *et al.* (1994).

The signal may be a consensus of proteins as in our examples, but also properties of single sequences like hydrophobicity of proteins or GC-distribution of DNA sequences are valid input data. In general any property of a single sequence or a set of sequences that can be assigned as discrete values to individual positions is suitable as input for CONRAD. However, not all possible applications were tested. Therefore, it remains to be determined how useful the results will be for other properties.

Due to back tracing of region boundaries from filtered to unfiltered signal, the scale-space method always yields level crossings of the unfiltered signal as boundaries which are not biased by any filtering. Region boundaries depend only on the signal itself, not on the smoothing. This is important for the general applicability of the method. In contrast, changing the window in the GCG PlotSimilarity will affect the threshold crossing.

Scale-space filtering is a model of human perception and will usually yield results very similar to intuitive perceived regions as well as their proposed boundaries. However, this is achieved on a precise algorithmic basis, avoiding any arbitrary choices. Therefore, it should be applicable to a wide range of protein sequences. The signal may be a consensus of proteins as in our examples, but also properties of single sequences like hydrophobicity or secondary structure prediction values of proteins or GC-distribution of DNA sequences are valid input data. Agreement of our results with experimental data indicates that our method should be a useful tool for generating hypotheses for subsequent experimental analysis.

Acknowledgements

We thank Kerstin Quandt, Kornelie Frech, and Markus Michaelis for critically reading the manuscript. Part of this work was supported by the BMBF Verbundprojekt GENUS 413-4001-01 IB 306 D (Förderschwerpunkt Bioinformatik).

References

Almagro,J.C., Vargas-Madrazo,E., Zenteno-Cuevas, R, Hernandez-Mendiola,V. and Lara-Orchoa,F. (1995) VIR: A computational tool for analysis of immunoglobulin sequences. *BioSystems*, **35**, 25–32.
Babaud,J., Witkin,A.P., Baudin,M. and Duda,R.O. (1986) Uniqueness of the Gaussian Kernel for Scale-Space Filtering. *IEEE Trans. Patt. Anal. Machine Intell. PAMI-8*, 26–33.
Barker,W.C., George,D.G., Mewes,H.W, Pfeiffer,F. and Tsugita,A. (1994) The PIR International Databases. *Nucleic Acids Res.*, **21**, 3089–3092.
Blinn,J.F. (1993) What's the Deal with the DCT? *IEEE Computer Graphics & Applications A13*, 78–83.
Chothia,C., Lesk,A M, Levitt,M., Amit,A.G., Mariuzza,R.A, Phillips,S.E.V. and Poljak,R.J. (1986) The predicted structure of

immunoglobulin D1.3 and its comparison with the crystal structure. *Science*, **233**, 755–758.

GCG, Genetics Computer Group (1994) *Program Manual for the Wisconsin Package, Version 8*. 575 Science Drive, Madison, WI.

Kabat,E.A. and Wu,T.T. (1970) Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N.Y. Acad. Sci.*, **190**, 382–393.

Kabat,E.A., Wu,T.T., Perry,H.M., Foeller,C and Gottesman,K.S. (1991) *Sequences of proteins of immunological interest*. Fifth edition of the Database of the US Department of Health and Human Services, Bethesda, MD.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Leonard,C.K., Spellman,M.W., Riddle,L., Harris,R.J., Thomas,J.N. and Gregory,T.J. (1990) Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese Hamster Ovary cells. J Biol. Chem., **265**, 10373–10382.

Lindeberg,T.P. (1990) Scale space for discrete signals. *IEEE Trans. Patt. Anal. Machine Intell. PAMI-12*, 234–254.

Lindeberg,T.P. (1993) Effective scale: A Natural Unit for Measuring Scale-Space Lifetime. *IEEE Trans. Patt. Anal. Machine Intell. PAMI-15*, 1068–1074.

Liu,Z.-Q. and Rangayyan,R.M. (1991) Directional analysis of images in scale space. *IEEE Trans. Patt. Anal. Machine Intell. PAMI-11*, 1185–1192.

Myers,G., Wain-Hobson,S., Henderson,L.E., Korber,B., Jeang,K.-T. and Pavlakis,G.N. (1994) *Human Retroviruses and AIDS*. Los Alamos National Lab., Los Alamos, NM.

Modrow,S., Hahn,B.H., Shaw,G.M., Gallo,R.C., Wong-Staal,F. and Wolf,H. (1987) Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: Prediction of antigenic epitopes in conserved and variable regions. *J. Virol.*, **61**, 570–578.

Moore,J.P., Sattentau,Q.J., Wyatt,R. and Sodroski,J. (1994) Probing the structure of the surface glycoprotein gp 120 of human immunodeficiency virus type 1 with a panel of monoclonal antibodies. *J. Virol.*, **68**, 469–484.

Schiffer,M., Mu,Z.B. and Clang,C.H. (1991) *Brookhaven Protein Database: entry 1BJL*

Witkin,A.P. (1983) Scale-Space Filtering. In *Proc. It. Con. Artificial Intell.* Karlsruhe.

Yuille,A.L., Poggio,T.A. (1986). Scaling theorems for zero crossings *IEEE Trans. Patt. Anal. Machine Intell. PAMI-8*, 15–25 .

Received on December 19, 1995; revised and accepted on May 22, 1996