Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 1638–1644
Bioorganic & Medicinal Chemistry

A simple, yet highly accurate, QSAR model captures the complement inhibitory activity of compstatin

Chandrika Mulakala,[a] John D. Lambris[b] and Yiannis Kaznessis[a,*]

[a]*Department of Chemical Engineering and Materials Science, and the Digital Technology Center, University of Minnesota, Minneapolis, MN 55455, USA*

[b]*Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA*

Received 18 September 2006; revised 27 November 2006; accepted 11 December 2006
Available online 13 December 2006

Abstract—Compstatin is a 13-residue cyclic peptide inhibitor of complement activation that was originally identified through phage-mediated presentation of a peptide library to C3b. Recent efforts to improve its activity have led to a rich dataset of complement analogs, with the most active analog being ~260 times more active than the parent compstatin. In the present work, a highly transparent quantitative structure–activity relationship model ($R^2_{adj} = 0.89$) with four parameters is presented that captures important physico-chemical and geometrical properties of the analog molecules with regard to activity. The number of aromatic bonds and hydrophobicity of the fourth residue of compstatin correlated strongly with activity. Also important were the hydrophobic patch size near the disulfide bond and the solvent-accessible surface area occupied by nitrogen atoms of basic amino acid residues.

1. Introduction

Compstatin is a 13-residue cyclic peptide inhibitor of complement activation that binds to C3, the third protein of complement.[1] The complement system consists of a large number of plasma and cell-surface proteins that are involved in host immune defense. However, uncontrolled complement activation can lead to rapid tissue damage, and its inappropriate activation has been implicated in a wide range of clinical conditions, such as various autoimmune diseases,[2] rejection following xenotransplantation,[3,4] Alzheimer's disease,[5,6] stroke,[7] and heart attack,[8] to name a few. C3, by virtue of its role as the connecting link between the classical, alternative, and lectin-mediated pathways of complement activation, is considered a suitable target for the development of complement inhibitors. Compstatin itself was discovered through the phage-mediated presentation of a peptide library to C3b.[1] Subsequent to its discovery, a multi-pronged approach involving a combination of experimental and computational techniques was employed to explore its inhibitory action, with the goal of improving its activity and thus its suitability as a therapeutic agent.[9] Its potential as a therapeutic has already been tested and demonstrated in several in vitro, in vivo, and ex vivo models.[1,3,4,10–12]

The sequence of compstatin is I[CVVQDWGHHRC]T-NH_2, with a disulfide bond between Cys2 and Cys12. A 2D NMR solution structure of a major conformation of compstatin reveals two important features: a type I β-turn involving Gln5-Asp6-Trp7-Gly8 and a hydrophobic patch that includes the disulfide bond.[13] Ala scan analogs of the cyclic part of compstatin identified Val-3 and the four β-turn residues as crucial for its inhibitory activity.[13] In a binding kinetics study using surface plasmon resonance, compstatin was found to bind C3 and its fragments C3b and C3c, but not to C3d.[10] In the same study, N-acetylation of compstatin was found to be necessary for improving activity, since in vitro studies indicated that a major biotransformation pathway involved the removal of Ile1. Interestingly, a side-effect of N-acetylation was a threefold increase in compstatin activity, which was hypothesized to be due to the extension of its hydrophobic patch by charge neutralization of the N-terminus.[14]

Rational design through computational and mutational analysis,[9] most recently with the incorporation of

Keywords: QSAR; Structure–activity relationships; Compstatin; C3; Complement.
* Corresponding author. Tel.: +1 612 624 4197; fax: +1 612 626 7246; e-mail: yiannis@cems.umn.edu

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.12.015

unnatural amino acids, led to a ~260-fold improvement in inhibitory activity, bringing the IC_{50} to 205 nM.[15] Despite these excellent advances, however, the economics of biotherapeutic development require a further decrease in the IC_{50}, preferably to a few nanomolar, thus leaving room for further improvement in its activity.

All the mutational analysis studies of compstatin outlined thus far have created a significant dataset of structure–activity data, opening up the potential for identifying physico-chemical or geometrical features of the compstatin analogs that are statistically correlated with the observed activity. This approach is generally referred to as quantitative structure–activity relationship (QSAR) determination. In this project, a QSAR study was conducted on all known active compstatin analogs to identify important molecular properties that relate to their activity. While a solution structure for compstatin exists,[13] as do X-ray crystallographic structures of C3 and C3c,[16] no structures have yet been reported for the compstatin-C3 complex. For the purpose of this study, an active analog was defined as one with an $IC_{50} < 100\ \mu M$ ($IC_{50,Rel} > 0.12$), which led to a dataset of 53 compstatin analogs (Table 1). Structures of the compstatin analogs were generated through in silico mutagenesis of the solution structure of compstatin,[13] and these structures were then used to determine 2D and 3D geometrical descriptors for QSAR analysis.

2. Results

The whole set of available compstatin analogs with their activities is shown in Table 1. The common logarithm of the IC_{50} relative to that of the parent compstatin ($IC_{50,Rel}$) was used to drive the statistical fit. An initial approach using all the 2D and 3D molecular descriptors available in MOE did not yield any model with significant statistical correlation (i.e., $R^2 > 0.85$). This result prompted us to take a different systematic, knowledge-based approach. This approach led to a much better model, with an R^2_{adj} of 0.89 and only four dependent variables, which is described in the following sections.

A careful examination of the available data led us to the fact that the most important modifications were those involving residues 4 and 7. In order to isolate the physico-chemical characteristics of these residues that lead to enhanced activity, one would need a subset of compstatin analogs with modifications to either residue 4 or residue 7 alone, while the rest of the molecule remained the same. Fortunately, a subset of 12 analogs is available that have mutations on residue 4 alone, along with an H9A mutation (AcCompNH$_2$_V4X/H9A series) (Table 1). A similar dataset for residue 7 does not exist. Also significant is the fact that the IC_{50} values for the AcCompNH$_2$_V4X/H9A subset of compstatin analogs span a wide range, from an $IC_{50,Rel}$ of 1.13 for AcCompNH$_2$_V4(Cha)/H9A to an $IC_{50,Rel}$ of 261 for AcCompNH$_2$_V4(1MeW)/H9A, outlining the importance of residue 4 for compstatin activity.

In order to isolate the physico-chemical characteristics of residue 4 that relate to their activity, the side chains alone of residue 4 for the AcCompNH$_2$_V4X/H9A subset were modeled, and a QSAR analysis was performed using molecular descriptors available in MOE. By isolating the side-chain structures, we assumed that these mutations did not cause significant changes in the backbone structure; this assumption is reasonable, since all the mutations to residue 4 in this data subset are fairly bulky. We found that the activity of these analogs correlated greatly ($R = 0.89$) with the number of aromatic bonds in the side chain of residue 4 (b_ar_4). When the dataset was expanded to include all the compstatin analogs, the correlation coefficient remained fairly high ($R = 0.88$); b_ar_4 was therefore retained as one of the parameters in the regression analysis.

Another pattern that was evident in the AcCompNH$_2$_V4X/H9A series was that the presence of polar side groups in residue 4 led to reduced activity. Therefore, when we had accounted for the polarity through a simple integer value corresponding to the number of polar atoms in the residue 4 side chain (a hydrophobic substitution was penalized by subtracting 1 from the total of polar side-group atoms in AcCompNH$_2$_V4(1MeW)/H9A and AcCompNH$_2$_V4(5MeW)/H9A), the R^2_{adj} increased from 0.77 to 0.90, thus establishing the negative effect of these polar atoms on the activity. The resultant QSAR model (Fig. 1) is as follows:

$$\ln[IC_{50,Rel}] = 0.1621\text{b\_ar\_4} - 0.2121\text{polar\_4} + 0.4093$$

$$n = 12 \quad R^2_{adj} = 0.90 \quad Q^2 = 0.81 \quad s = 0.2184$$

$$s_{cv} = 0.2799.$$

Compstatin has a hydrophobic patch close to Cys2 and Cys12, whose importance for activity has been discussed earlier.[17] We realized that the hydrophobic patch was not being represented by the properties calculated by MOE, since MOE only calculates properties for the whole peptide. The SASA for the hydrophobic patch was therefore calculated using Pymol.[21] When this parameter was included in the least squares fit for the whole molecule, the R^2_{adj} increased from 0.79 to 0.85. In addition, calculating the SASA occupied by the nitrogen atoms of basic residues improved the statistical fit from 0.85 to 0.88. The polarity of residue 4 did not contribute significantly to the overall dataset, as it did for the AcCompNH$_2$_V4X/H9A set of analogs. Including polar_4 increased the R^2_{adj} marginally from 0.88 to 0.89.

Our QSAR equations for the entire compstatin analog set with three and four (Fig. 2) parameters are:

$$\log_{10}[IC_{50,Rel}] = 0.1457\text{b\_ar\_4} + 0.0032\text{hyd\_patch\_surf} + 0.0051\text{base\_N\_surf} - 1.7538$$

Table 1. Activity dataset of compstatin analogs with QSAR descriptors

Compstatin analog[a]	$IC_{50,Rel}$[b]	log[$IC_{50,Rel}$]	b_ar_4	polar_4	hyd_patch_surf	base_N_surf	Reference
CompCC	0.364	−0.439	0	0	180.203	148.941	1
CompCC_H10A	0.162	−0.790	0	0	190.555	125.712	1
CompCC_H9A	0.800	−0.097	0	0	233.072	126.369	1
CompCC_R11A	0.171	−0.767	0	0	181.829	56.140	1
$AcCompNH_2$	3.000	0.477	0	0	381.061	139.970	2
$CompNH_2$	1.000	0.000	0	0	307.243	149.116	2
$CompNH_2$_del1I	0.480	−0.319	0	0	220.977	148.216	2
$CompCCNH_2$	0.364	−0.439	0	0	220.427	148.404	2
$AcCompNH_2$_H9A[c]	4.138	0.617	0	0	434.507	121.154	3
$AcCompNH_2$_H9A/R11A	1.212	0.084	0	0	434.801	27.155	3
$AcCompNH_2$_Q5N	2.857	0.456	0	0	381.507	141.567	3
$AcCompNH_2$_Q5N/R1 1A	0.200	−0.699	0	0	382.986	50.483	3
$AcCompNH_2$_R11S	0.480	−0.319	0	0	380.182	48.083	3
$AcCompNH_2$_T13I	3.750	0.574	0	0	489.308	141.431	3
$AcCompNH_2$_V3L	1.200	0.079	0	0	413.414	142.148	3
$AcCompNH_2$_V3L/Q5N	1.446	0.160	0	0	423.487	141.912	3
$CompNH_2$_R11K	0.594	−0.226	0	0	308.380	104.631	3
$AcCompNH_2$_I1D	0.545	−0.264	0	0	169.366	139.984	4
$AcCompNH_2$_I1L/H9W/T13G	4.138	0.617	0	0	459.596	119.493	4
$AcCompNH_2$_I1M/V4H/H9G/T13F	0.137	−0.863	6	3	381.409	145.543	4
$AcCompNH_2$_I1R	1.500	0.176	0	0	172.461	254.923	4
$AcCompNH_2$_I1S/V4F/H9R/H10L/R11A/T13P	0.500	−0.301	6	0	192.819	61.818	4
$AcCompNH_2$_V4A/H9A/T13I	3.000	0.477	0	0	517.555	124.942	4
AcComp _I1dI/V4W/H9A	2.320	0.365	6	2	411.084	121.604	5
AcComp_V4(1Nal)/H9A	29.778	1.474	11	0	382.672	112.694	5
AcComp_V4(2Igl)/H9A	35.773	1.554	6	0	382.329	118.480	5
AcComp_V4(2Nal)/H9A	38.286	1.583	11	0	385.721	110.487	5
AcComp_V4(Bpa)/H9A	48.727	1.688	12	1	385.736	113.539	5
AcComp_V4(Bta)/H9A	67.000	1.826	10	1	463.216	123.358	5
AcComp_V4W/H9A/T13dT	19.852	1.298	6	2	467.695	119.945	5
$AcCompNH_2$ V4(2Igl)/H9A[c]	38.286	1.583	6	0	382.329	114.123	5
$AcCompNH_2$_V4(Bpa)/H9A[c]	89.333	1.951	12	1	385.736	102.082	5
$AcCompNH_2$_V4(Bta)/H9A[c]	67.000	1.826	10	1	463.216	120.417	5
$AcCompNH_2$_V4(Cha)/H9A[c]	1.138	0.056	0	0	500.490	117.392	5
$AcCompNH_2$_V4(Dht)/H9A[c]	5.255	0.721	6	2	467.454	110.627	5
$AcCompNH_2$_V4F	5.255	0.721	6	0	407.153	133.571	5
$AcCompNH_2$_V4H	5.105	0.708	6	3	303.076	162.739	5
$AcCompNH_2$_V4S	1.053	0.022	0	2	303.922	139.984	5
$AcCompNH_2$_V4T	0.785	−0.105	0	2	303.991	140.139	5
$AcCompNH_2$_V4W	24.364	1.387	10	2	420.737	133.463	5
$AcCompNH_2$_V4W/H9(Abu)	35.733	1.553	10	2	434.329	121.271	5
$AcCompNH_2$_V4W/H9W	17.290	1.238	10	2	513.896	114.017	5
$AcCompNH_2$_V4Y/H9A[c]	14.889	1.173	6	2	432.622	116.440	5
$AcCompNH_2$_V4(5-OH-W)/W7(5-OH-W)/H9A	1.624	0.211	10	4	441.391	124.050	6
$AcCompNH_2$_V4(6fW)/W7(6fW)/H9A	124.651	2.096	10	3	446.562	124.0	6
$AcCompNH_2$_V4(7-aza-W)/W7(7-aza-W)/H9A	0.439	−0.358	10	3	471.141	124.050	6
$AcCompNH_2$_V4W/H9A[c]	44.667	1.650	10	2	462.754	121.271	6
$AcCompNH_2$_V4(1MeW)/H9A[c]	261.463	2.417	10	−1	479.338	113.543	7
$AcCompNH_2$_V4(1MeW)/W7(5fW)/H9A	261.463	2.417	10	−1	478.715	113.543	7
$AcCompNH_2$_V4(2Nal)/H9A[c]	98.350	1.993	11	0	385.721	106.826	7
$AcCompNH_2$_V4(5fW)/H9A[c]	30.805	1.489	10	3	442.284	121.271	7
$AcCompNH_2$_V4(5MeW)/H9A[c]	61.609	1.790	10	2	417.040	117.708	7
$AcCompNH_2$_V4W/W7(5fW)/H9A	120.279	2.080	10	2	463.721	121.271	7

Refs. 1: Morikis et al. *Protein Sci.* **1998**, *7*, 619–627; 2: Sahu et al. *J. Immunol.* **2000**, *165*, 2491–2499; 3: Morikis et al. *J. Biol. Chem.* **2002**, *277*, 14942–14953; 4: Soulika et al. *J. Immunol.* **2003**, *170*, 1881–1890; 5: Mallik et al. *J. Biol. Chem.* **2005**, *48*, 274–286; 6: Katragadda; Lambris, *Protein Expr. Purif.* **2006**, *47*, 289–195; 7: Katragadda et al. *J. Med. Chem.* **2006**, *49*, 4616–4622.

[a] Comp, compstatin; CompCC, I1,13T deletion analogs of compstatin; Ac, N-terminal acetylation; NH_2, C-terminal amidation; del, deletion; d, D-amino acid.

[b] IC_{50} relative to parent compstatin.

[c] AcCompNH2_V4X/H9A subset, where X is any amino acid, natural or unnatural.

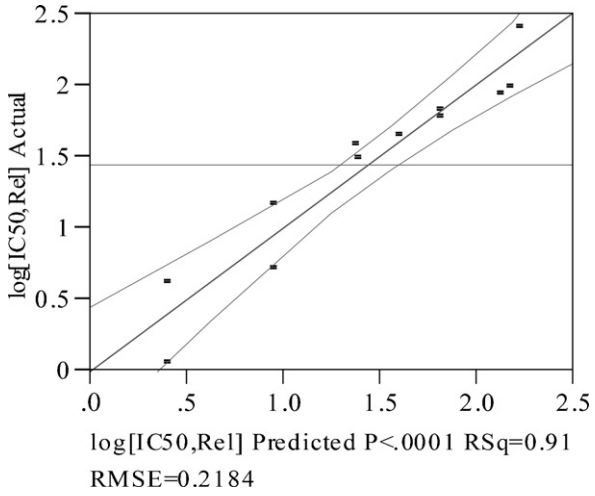

Figure 1. Plot of actual versus predicted log[$IC_{50,Rel}$] for the AcCompNH$_2$_V4X/H9A set of compstatin analogs along with 95% confidence curves generated using JMP.[22]

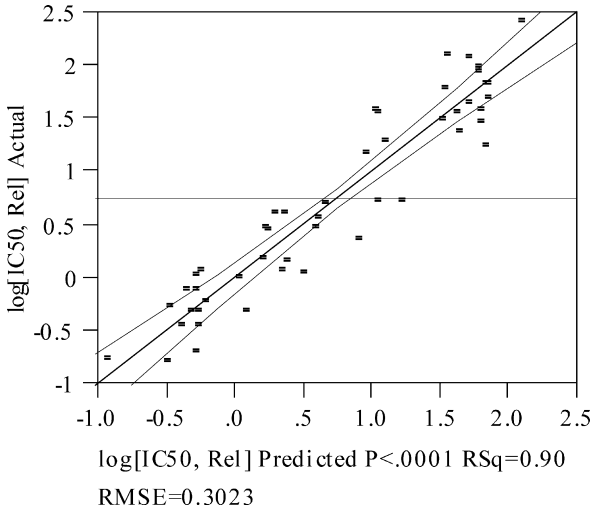

Figure 2. Plot of actual versus predicted log[$IC_{50,Rel}$] for the complete dataset using the four-parameter model along with 95% confidence curves generated using JMP.[22]

$$n = 50 \quad R^2_{adj} = 0.88 \quad Q^2 = 0.87 \quad s = 0.3237$$
$$s_{cv} = 0.3323$$

$$\log_{10}[IC_{50,Rel}] = 0.1573\text{b\_ar\_4} + 0.0034\text{hyd\_patch\_surf} + 0.0058\text{base\_N\_surf} - 0.1266\text{polar\_4} - 1.8572$$

$$n = 50 \quad R^2_{adj} = 0.89 \quad Q^2 = 0.88 \quad s = 0.3023$$
$$s_{cv} = 0.3167.$$

The models reported above for the full dataset had three outliers, AcCompNH2_I1M/V4H/H9G/T13F, AcCompNH$_2$_V4(5-OH-W)/W7(5-OH-W)/H9A, and AcCompNH2_V4(7-aza-W)/W7(7-aza-W)/H9A. Inclusion of the outliers resulted in a QSAR model with an R^2_{adj} of 0.76 with all four parameters.

3. Discussion

Of the three descriptors that are included in the QSAR equation, b_ar_4 seemed to be the most important for activity. The b_ar_4 parameter captured two important characteristics of the side chains—their "flatness", as well as the area they occupy, since the number of aromatic bonds would equal the number of sides to the flat polygonal structures formed by the side chains. Hydrophobicity seemed to be very important as well, since the inclusion of a simplistic account of their polarity improved the statistical fit greatly (R^2_{adj} changed from 0.77 to 0.90 for the AcCompNH$_2$_V4X/H9A subset). The effect of the V4W mutation on the area of the hydrophobic patch is shown in Figure 3. The bulky tryptophan seems to consolidate and extend the hydrophobic patch that most likely forms a part of the compstatin-C3 binding interface. If this were indeed the case, the presence of polar groups in this patch would reduce their hydrophobicity, which seems to be measured effectively by the polar_4 parameter. The possible interaction of W4 with an aromatic residue of C3 through hydrophobic stacking interactions, or even the presence of a cation-π interaction, has been previously discussed[14] and cannot be ruled out.

While both the hydrophobic patch[17] and residue 4[18] have been previously thought to be important, another feature of the parent compstatin, the presence of three contiguous basic amino acids H9, H10, and R11, has not been explicitly discussed thus far. One can see that while the H9A analog showed improved activity, mutations to R11 generally seemed to reduce activity. This relationship led us to evaluate base_N_surf, which improved the statistical fit further. The significance of the base_N_surf to the molecule's activity, however, is relatively hard to interpret. The overall positive charge of the molecule is not the primary factor here, since an R11K mutation led to a decrease in the activity. Thus, the number of hydrogen bond donors may be significant and, more specifically, Arg11 may be particularly important for activity.

The reduced impact of polar_4 on the final model may be explained by the fact that it assumed greater significance only when bulky side chains were present, since the addition of polar_4 improved the statistical fit for the AcCompNH$_2$-V4X/H9A subset of analogs but not for the whole dataset. This result highlights an important limitation of the QSAR method: that the choice of the dataset greatly affects the final model. Therefore, even though it is generally accepted that the principle of parsimony (Occam's razor) should be adopted for parameter selection in QSAR analyses, we believe that polar_4 should be retained as a parameter in the QSAR model because its negative effect on activity seems to be important.

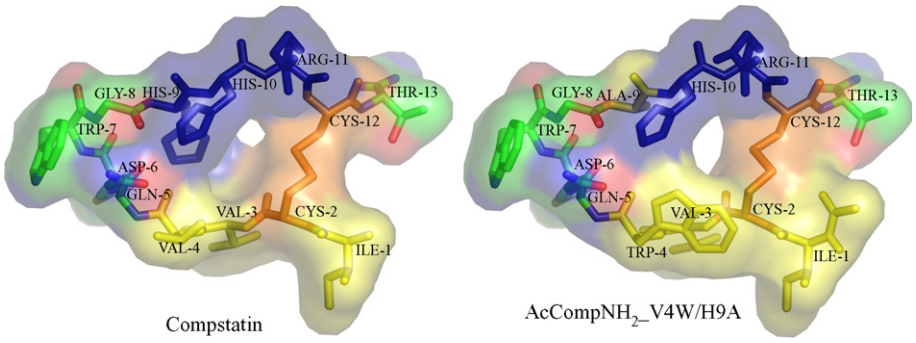

Figure 3. Structures of compstatin and the modeled AcCompNH$_2$_V4W/H9A show how V4W and related mutations increase the surface of the hydrophobic patch (yellow) near the Cys2-Cys12 disulfide bond (orange). Basic residues are shown in dark blue, and the rest of the structure is rendered in CPK colors.

It was very surprising that we were able to obtain a QSAR model with very high statistical significance in the absence of a ligand–receptor complex and through the use of just four structural parameters with three outliers. An explanation for this success may be related to the observation of Cronin and Schultz[19] that the QSAR can only be as good as the quality of the data itself. Here we had the ideal case for the dataset, with all of the data having been generated in the same laboratory, by the use of the same protocol. Also, the absence of a ligand–receptor complex does not seem to have been a deterrent in our QSAR model because of the nature of our choice of descriptors. Three of the four properties in the QSAR model obtained, the aromaticity and polarity of residue 4 and the basicity analogs, were more or less independent of the analog's tertiary structure. Also, whatever the tertiary structure of the analog in the analog-C3 complex, it seems that it results in the formation of a contiguous hydrophobic patch similar to the one observed in compstatin's solution structure. The significance of a highly hydrophobic mutation in residue 4[18] and the hydrophobic patch[17] has been previously discussed, and this work reaffirms their importance.

Cronin and Schultz[19] also talk about the importance of avoiding collinear parameters in the modeling, which can lead to very high R^2 and introduce instability in the QSARs. They therefore suggest that a correlation matrix be obtained for QSAR model parameters, and to avoid collinearity, the correlation coefficients between them should be low. While there is no agreement in the QSAR community on an acceptable level of collinearity, they suggest that it should be much lower than the statistical fit of the model itself. For our model, all three parameters had a very low correlation coefficient (<0.44, Table 2). The highest collinearity was observed between b_ar_4 and hyd_patch_surf. A certain degree of correlation between the two was expected, since a significant portion of the hydrophobic area of the hydrophobic patch is contributed by the hydrophobic mutations in residue 4; however, the hyd_patch_surf parameter can and does capture the effect of mutations in other residues adjacent to the disulfide bond as well. Cronin and Schultz[19] have also suggested that instead of the R^2 value, one should report the R^2_{adj}, which is what we have reported throughout this paper.

An interesting aspect of the base_N_surf parameter can be observed in the correlation matrix (Table 2): its correlation with the $IC_{50,Rel}$ was very low. This low correlation of base_N_surf with activity (Table 2) seems to reflect the fact that a significant fraction of the mutations in a basic residue were those of H9A, which resulted in a 4-fold increase in activity. Here, the effect of increased hydrophobicity seems to dominate over the reduced basicity. The base_N_surf parameter, nevertheless, effectively captured the negative effect of mutations in R11.

The three outliers in the resulting model outline the limitations of QSAR analyses. The fact that there is no significant dataset for quantifying the effect of changes to residue 7 suggests that those effects are underrepresented in the dataset and would explain why the activity of AcCompNH$_2$_V4(5-OH-W)/W7(5-OH-W)/H9A and AcCompNH$_2$_V4(7-aza-W)/W7(7-aza-W)/H9A could not be predicted by our QSAR model. In addition, b_ar_4 and polar_4 may be too simplistic to completely capture the steric and hydrophobic character of the mutations to residue 4. The

Table 2. Correlation matrix of QSAR variables used in least-squares regression analysis

	b_ar_4	polar_4	hyd_patch_surf	base_N_surf	log[$IC_{50,Rel}$]
b_ar_4	1.0000	0.4329	0.4433	−0.1181	0.8801
polar_4	0.4329	1.0000	0.2489	0.1014	0.3147
hyd_patch_surf	0.4433	0.2489	1.0000	−0.2131	0.6408
base_N_surf	−0.1181	0.1014	−0.2131	1.0000	0.0226
log[$IC_{50,Rel}$]	0.8801	0.3147	0.6408	0.0226	1.0000

low activity of AcCompNH$_2$_I1M/V4H/H9G/T13F is probably due to the H9G mutation. Glycine residues are known to increase backbone flexibility, while our assumption for the QSAR analysis has been that the mutations do not greatly affect backbone flexibility of the analogs.

4. Conclusion

There are, in general, no specific guidelines for QSAR development. Our approach toward QSAR development in this project was to first intuitively derive hypotheses and then test them through statistical analysis. For example, for the current dataset, it had previously been hypothesized that the hydrophobic patch is an important contributor to compstatin's activity.[17] Also, mutations to residue 4 were hypothetically driven.[18] Our QSAR analysis helped not only to verify these hypotheses but also to quantify their relative effects. Once these effects have been quantified, subtle effects, such as the effect of the base_N_surf parameter, can then be deciphered. This hypothesis-driven approach not only helps to generate highly transparent QSAR models but can also be useful in strengthening and directing one's intuition, hopefully toward the development of analogs with higher activity. For statistical analyses, after all, are not amenable to extrapolation beyond the range of the properties in the dataset, whereas the goal of rational design is to improve activity, typically by several orders of magnitude.

5. Computational methods

Analogs of compstatin were generated using MOE[20] through in silico point mutations on the NMR solution structure of compstatin (PDB ID: 1A1P). Atoms within 4.5 Å of the mutated residue were minimized using the CHARMm force-field in MOE. Minimization was performed in such a way as to optimize the mutated side-chain conformations without causing any significant alterations in the backbone conformation. The initial QSAR analysis on the AcCompNH$_2$-V4X/H9A analog subset (Table 1) was performed using MOE. The side chains alone were synthesized in MOE, and then all of the 2D and 3D molecular descriptors available with the MOE QSAR module were systematically examined for correlation with activity. For QSAR of the full dataset, hyd_patch_surf and base_N_surf (Table 1) were computed using Pymol,[21] where hyd_patch_surf is the solvent accessible surface area in Å^2 (SASA) of the side-chain carbons of hydrophobic residues adjacent to the Cys2-Cys12 disulfide bond, and base_N_surf is the SASA occupied by nitrogen atoms of the basic residues. Based on the solution structure of compstatin, only hydrophobic amino acids of residues 1, 3, 4, 6, 9, and 13 were considered to be part of the hydrophobic patch. In both cases, hydrogens were removed before computation of SASA. The statistical analysis for generation of the QSAR models was carried out using JMP.[22] The leave-one-out cross validation co-efficient, Q^2, and the root mean square error of cross validation (s_{cv}) were computed using MATLAB.[23] The various other terms reported for the resulting least-squares regression equations for the QSARs are: the number of compounds, n, the adjusted correlation coefficient, R^2_{adj}, and the standard deviation (or the root mean square error), s.

Acknowledgments

We also thank Deborah McClellan for editorial assistance. This work was supported by NIH Grants GM 62134 and GM 069736.

References and notes

1. Sahu, A.; Kay, B. K.; Lambris, J. D. *J. Immunol.* **1996**, *157*, 884–891.
2. Kalli, K. R.; Hsu, P.; Fearon, D. T. *Springer Semin. Immun.* **1994**, *15*, 417–431.
3. Fiane, A. E.; Mollnes, T. E.; Videm, V.; Hovig, T.; Hogasen, K.; Mellbye, O. J.; Spruce, L.; Moore, W. T.; Sahu, A.; Lambris, J. D. *Xenotransplantation* **1999**, *6*, 52–65.
4. Fiane, A. E.; Mollnes, T. E.; Videm, V.; Hovig, T.; Hogasen, K.; Mellbye, O. J.; Spruce, L.; Moore, W. T.; Sahu, A.; Lambris, J. D. *Transplant. Proc.* **1999**, *31*, 934–935.
5. Rogers, J.; Cooper, N. R.; Webster, S.; Schultz, J.; McGeer, P. L.; Styren, S. D.; Civin, W. H.; Brachova, L.; Bradt, B.; Ward, P., et al. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10016–10020.
6. Bradt, B. M.; Kolb, W. P.; Cooper, N. R. *J. Exp. Med.* **1998**, *188*, 431–438.
7. Vasthare, U. S.; Barone, F. C.; Sarau, H. M.; Rosenwasser, R. H.; DiMartino, M.; Young, W. F.; Tuma, R. F. *Brain Res. Bull.* **1998**, *45*, 413–419.
8. Kilgore, K. S.; Friedrichs, G. S.; Homeister, J. W.; Lucchesi, B. R. *Cardiovasc Res.* **1994**, *28*, 437–444.
9. Morikis, D.; Soulika, A. M.; Mallik, B.; Klepeis, J. L.; Floudas, C. A.; Lambris, J. D. *Biochem. Soc. Trans.* **2004**, *32*, 28–32.
10. Sahu, A.; Soulika, A. M.; Morikis, D.; Spruce, L.; Moore, W. T.; Lambris, J. D. *J. Immunol.* **2000**, *165*, 2491–2499.
11. Nilsson, B.; Larsson, R.; Hong, J.; Elgue, G.; Ekdahl, K. N.; Sahu, A.; Lambris, J. D. *Blood* **1998**, *92*, 1661–1667.
12. Soulika, A. M.; Khan, M. M.; Hattori, T.; Bowen, F. W.; Richardson, B. A.; Hack, C. E.; Sahu, A.; Edmunds, L. H.; Lambris, J. D. *Clin. Immunol.* **2000**, *96*, 212–221.
13. Morikis, D.; Assa-Munt, N.; Sahu, A.; Lambris, J. D. *Protein Sci.* **1998**, *7*, 619–627.
14. Morikis, D.; Roy, M.; Sahu, A.; Troganis, A.; Jennings, P. A.; Tsokos, G. C.; Lambris, J. D. *J. Biol. Chem.* **2002**, *277*, 14942–14953.
15. Katragadda, M.; Magotti, P.; Sfyroera, G.; Lambris, J. D. *J. Med. Chem.* **2006**, *49*, 4616–4622.
16. Janssen, B. J.; Huizinga, E. G.; Raaijmakers, H. C.; Roos, A.; Daha, M. R.; Nilsson-Ekdahl, K.; Nilsson, B.; Gros, P. *Nature* **2005**, *437*, 505–511.
17. Soulika, A. M.; Morikis, D.; Sarrias, M. R.; Roy, M.; Spruce, L. A.; Sahu, A.; Lambris, J. D. *J. Immunol.* **2003**, *171*, 1881–1890.

18. Mallik, B.; Katragadda, M.; Spruce, L. A.; Carafides, C.; Tsokos, C. G.; Morikis, D.; Lambris, J. D. *J. Med. Chem.* **2005**, *48*, 274–286.
19. Cronin, M. T. D.; Schultz, T. W. *J. Mol. Struc-Theochem.* **2003**, *622*, 39–51.
20. MOE: Molecular Operating Environment; Chemical Computing Group: Montreal, Canada, 2005. http://www.chemcomp.com.
21. DeLano, W.L.; The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, USA, 2002. http://www.pymol.org.
22. JMP statistics software; SAS Institute: Cary, NC, USA, 1989. http://www.jmp.com.
23. MATLAB; The MathWorks, Inc.: Natick, MA, USA, 2006. http://www.mathworks.com.