Why F1(prob) is incorrect?

F1 is a well-established metric to evaluate the performance of a classifier or a predictor. The issue with F1 is that it can only evaluate the ranking or classification capability of a predictor, but not the quality of the probability values assigned to predicted contacts. Intuitively, a probability assignment is good if it has the following properties: 1) it can help users separate a good prediction from a bad one. That is, for a hard target, most predicted probability values shall have a small value; while for an easy target, there shall be many more large probability values among the top predictions; and 2) it can help users separate contacts from non-contacts. That is, on average a true contact (especially those top ranked) shall have a much larger predicted probability value than a non-contact. The above properties are desirable since they can facilitate users to guess if a prediction is of low- or high-quality and to select top predicted contacts for folding simulation (in the absence of native structures).

To evaluate the quality of the predicted probability values, the CASP12 contact prediction assessor introduced a new performance metric called probability-weighted F1, abbreviated as F1(prob). Generalized from $\mathrm{F1} = \frac{2*\mathrm{precision}*\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$, F1(prob) is defined as $\frac{2*\mathrm{precision(prob)}*\mathrm{recall(prob)}}{\mathrm{precision(prob)}+\mathrm{recall(prob)}}$ where $\mathrm{precision(prob)} = \frac{probability\_sum}{\#top\ predictions}$, $\mathrm{recall(prob)} = \frac{probability\_sum}{\#contacts}$ and $probability\_sum$ is the sum of the probability values assigned to all true contacts among the top predictions. Intuitively, one probability assignment with a larger F1(prob) shall be better than another with a smaller F1(prob) in separating a good prediction from a bad one and separating contacts from non-contacts. Although seems to be appealing, it turns out that F1(prob) is a wrong metric mainly because it does not penalize a large probability value assigned to a residue pair not forming a contact. Below please see a more detailed argument.

Let L denote the length of a protein under prediction and N the number of residue pairs. Let $P = \{P_1, P_2, \ldots, P_N\}$ denote the probability values assigned by one predictor to the N residue pairs where $P_1 > P_2 \ldots > P_N$. Let $F1(prob)$ (P) denote the $F1(prob)$ of this prediction. Supposing that this protein is a hard target with very few sequence homologs, this prediction shall have a low quality and most of the predicted probability values shall be smaller than 0.5. That is, many of the top L predicted contacts are actually wrong.

Now we show that there exist numerous probability assignments, denoted by Q, such that 1) Q has the same ranking order as P, i.e., P and Q have the same F1 score; 2) Q has a better F1($prob$), i.e., $F1(prob)(Q) > F1(prob)$ (P) ; 3) Q may

mislead a user to believe that the underlying prediction is of high quality although it is indeed of low quality; and 4) Q may be worse than P in separating contacts from non-contacts.

A trivial case. We can obtain Q by assigning the probability values of the top L residue pairs to a value close to 1 without changing their ranking order and assigning the remaining N-L probability values to 0 or a small constant (e.g., 0.01). Since the ranking order of the residue pairs is not changed, Q has the same F1 as P, but a larger F1(prob) than P. However, since the top L probability values are close to 1 and much larger than the remaining ones, when presented a probability assignment Q instead of P, a user may think that the underlying prediction is really good and the top L predicted contacts are very likely to be true positives, which contradicts with the fact that the underlying prediction is actually bad and many of the top L predicted contacts are false positives.

A nontrivial case. We may set Q to $\{\sqrt{P_1}, \sqrt{P_2}, \sqrt{P_3}, \ldots, \sqrt{P_N}\}$. In fact, for any positive $\alpha \leq 1/2$, we can set Q to P^{α} to have the same effect. In addition, we can just change the top L probability values and keep the remaining N-L probability values unchanged. We can also do linear interpolation between P and P^{α} to generate Q. It is easy to show that Q has the same F1 as P, but a larger F1(prob) than P. However, since many small probability values in P become large in Q (e.g., 0.25 becomes 0.5), when presented a probability assignment Q instead of P, a user may think that the underlying prediction is of high quality and many top predicted contacts are correct, which again contradicts with the fact that the underlying prediction actually has bad quality.

In summary, F1(prob) favors those predictors which would like to assign a larger probability value to a predicted contact no matter whether it is correct or not. F1(prob) cannot be used to evaluate the quality of the predicted probability values since one probability assignment with a larger F1(prob) is not better (sometimes even worse) than another with a smaller F1(prob) in separating contacts from non-contacts and in separating good predictions from bad. If F1(prob) is enforced, every server will simply just assign all the predicted contacts with probability 1.0, which is useless, to maximize their F1(prob).

P.S. A similar argument can also be used to show that the current method used by the assessor to filter predictions by probability>0.5 is also wrong.